DOCUMENT RESUME

ED 194 605 TM 800 736

TITLE	Issues in the Analysis and Analysis of Change of National Assessment Data.
INSTITUTION	Education Commission of the States, Denver, Colo. National Assessment of Educational Progress.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
REPORT NO	NAEP-12-IP-57
PUB DATE	Sep 80
GRANT	NIE-G-80-0003
NOTE	108p.
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	*Data Analysis; *Educational Assessment; Elementary Secondary Education; Measurement Techniques; *National Surveys; *Outcomes of Education; Research Design; *Trend Analysis; Young Adults
IDENTIFIERS	*National Assessment of Educational Progress

ABSTRACT

Approaches to data analysis used by the National Assessment of Educational Progress (NAEP) and possible alternative methodologies are reviewed. Since various components of the NAEP model affect analysis, key elements of that model, including procedures for developing objectives and exercises, drawing the sample, collecting data and reporting results, are briefly described. Performance on sets of items is summarized by using the average of percentages of acceptable responses to the items in a set. Methods for computing the statistics commonly used by NAEP and the standard errors of those statistics are described. The paper presents answers to questions frequently raised regarding NAEP analytical procedures, including such topics as the use of NAEP data to establish causal relationships, use of data adjustments, possibility of increasing the number of background questions, potential for setting performance standards, provision of different data breakdowns and new methods of analysis. For the 1979-80 reading assessment, data will be reported by achievement class, a new reporting variable. Appendices define reporting variables, present forms used to collect background information, describe population proportion smoothing methods, provide information about the problem of nonresponse in assessment samples, and describe methods of accounting for sample design with design effects. (Author/GK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED194605

TM

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ISSUES IN THE ANALYSIS AND ANALYSIS OF CHANGE OF NATIONAL ASSESSMENT DATA

No. 12-IP-57

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
T. Pratt

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

by the Staff of the
National Assessment of Educational Progress

Education Commission of the States
Suite 700, 1860 Lincoln Street
Denver, Colorado 80295

September 1980

TM800736

The National Assessment of Educational Progress is funded by the National Institute of Education. It is under contract with the Education Commission of the States. It is the policy of the Education Commission of the States to take affirmative action to prevent discrimination in its policies, programs and employment practices.

The National Assessment of Educational Progress is an education research project mandated by Congress to collect and report data, over time, on the performance of young Americans in various learning areas. National Assessment makes available information on assessment procedures and materials to state and local education agencies and others.

The work upon which this publication is based was performed pursuant to Grant No. NIE-G-80-0003 of the National Institute of Education. It does not, however, necessarily reflect the views of that agency.

TABLE OF CONTENTS

PREFACE 1

INTRODUCTION 3

SECTION 1: OBJECTIVES DEVELOPMENT 6

SECTION 2: DEVELOPMENT OF EXERCISES 7

SECTION 3: PREPARATION OF ASSESSMENT MATERIALS 9

SECTION 4: SAMPLING 11

SECTION 5: DATA COLLECTION 13

SECTION 6: SCORING 15

SECTION 7: DATA ANALYSIS 16
- Measures of Achievement 16
  - Estimating Variability in Achievement Measures . 18
  - Controlling Systematic Errors 19
  - Reporting the Data 20
  - Summary Reports 20
  - Public-Use Data Tapes 21
  - Microfiche 21
  - User Services 21

SECTION 8: COMPUTATION OF MEASURES OF ACHIEVEMENT, CHANGES IN ACHIEVEMENT AND STANDARD ERRORS . 22
- Measures of Achievement 22
- Computation of Standard Errors 25

SECTION 9: SOME ANALYSIS ALTERNATIVES FOR DETERMINING CHANGE IN STUDENT PERFORMANCE 30
- Longitudinal Versus Cross-Sectional Studies 30
- Alternative Exercise Development and Release Strategies 31
- Latent Trait Models 33
- Other Approaches 34

SECTION 10: QUESTIONS FREQUENTLY RAISED CONCERNING NAEP ANALYSIS 35

SECTION 11: CONCLUSIONS AND RECOMMENDATIONS 45

DEC 4 1980

APPENDIX A: DEFINITION OF NATIONAL ASSESSMENT REPORTING GROUPS 46

APPENDIX B: FORMS USED TO OBTAIN BACKGROUND INFORMATION 50

APPENDIX C: SAMPLES OF RESPONDENT LEVEL BACKGROUND QUESTIONS RELATED TO SPECIFIC SUBJECT AREAS 65

APPENDIX D: ADJUSTMENT OF RESPONDENT WEIGHTS BY SMOOTHING TO REDUCE RANDOM VARIABILITY OF ESTIMATED POPULATION PROPORTIONS 81

APPENDIX E: NONRESPONSE IN ASSESSMENT SAMPLES 90

APPENDIX F: INFERENCE FROM SURVEY DATA AND ANALYSIS OF DATA FROM COMPLEX SAMPLE SURVEYS 93

APPENDIX G: ANALYSIS ADVISORY COMMITTEE (ANAC) 99

BIBLIOGRAPHY 101

PREFACE

The following paper is one of seven required by the National Institute of Education (NIE) grant to the Education Commission of the States for the National Assessment of Educational Progress (NAEP). The papers are intended to raise significant questions about NAEP functions and procedures in order to apprise the Assessment Policy Committee of its technical and administrative options as NAEP moves into a second decade of activity. According to the NIE Program Announcement, each paper should include "discussions about the significance of a national assessment of educational progress for (a) informing public opinion, (b) influencing educational and other social policies, (c) identifying specific educational and research needs, and (d) contributing to the goal of improving education." Thus, the papers combine broad theoretical considerations with very concrete discussions of potential options and their trade offs. All the papers are suggestive, rather than definitive--starting points for discussion rather than end points.

Two fundamental questions undergird these papers. First, if one were to design a national assessment program from scratch to serve the needs of the eighties and nineties, what would it look like? What theory would it draw upon? What constituents would it serve? How would it operate? And second, given that an assessment has been operating for ten years, how do answers to these questions match the current NAEP? How could the current NAEP be changed to accommodate needs and technology unforeseen when it was begun in the 1960s? As these questions weave in and out of the discussions of various issues, it will become clear that the answers one gives will probably depend upon one's assumptions about such things as the way society, education and science work, as well as one's familiarity with the theory, politics, history and operation of the current National Assessment.

There are many ways to implement the 1978 legislation mandating a national assessment. Some are undoubtedly better than others in certain respects; all have both advantages and drawbacks, complicating even the simplest comparisons among alternatives. In such a situation, it is imperative that advocates for various models of a national

assessment adopt an open, interdisciplinary approach to rethinking the assessment, listening carefully to each other and respecting professional differences of opinion.

Individuals charged with making policy decisions about NAEP--members of the Assessment Policy Committee and members of the U.S. Congress--will find that, in many instances, attempts to strengthen the Assessment in one regard will weaken it in some other regard. Given a certain level of funding, advances in some areas come at the expense of others. The National Assessment can neither be all things to all people nor a single, unchanging enterprise. It will always represent a combination of various tensions. The question for its future is not how do we eliminate those tensions? Rather it is: How do we harness those tensions creatively to keep the Assessment abreast of the times and continually relevant to the needs of American education?

INTRODUCTION

The purpose of this paper is to discuss the analysis of data collected by the National Assessment of Educational Progress (NAEP) and, in particular, appropriate approaches for analyzing change. In simplest terms, the analysis can be characterized as establishing baseline estimates of the percentages of young Americans possessing certain skills, knowledge, understandings and attitudes and producing estimates of changes in these percentages over time.

A distinguished panel of statisticians and measurement experts has been closely associated with the Assessment since its inception. Initially, this panel, the Analysis Advisory Committee (ANAC), was responsible for creating the basic design of the project. Through the years, ANAC has continued to guide the Assessment staff in technical matters. They have provided valuable direction in virtually all aspects of the Assessment and have developed and refined the present analysis techniques. The ANAC members from 1965 to the present are listed in Appendix G.

In addition, for the governance of National Assessment the Education Commission of the States as directed by the legislation authorizing the program has established the Assessment Policy Committee (APC). Members represent state and local education agencies, state legislatures, professional education associations, and the business community. A recently constituted subgroup of the APC has responsibility for formulating policy decisions regarding operations and analysis.

It should be emphasized that National Assessment is not designed to monitor changes or growth in achievement of individual students. Its purpose is to measure the impact of a changing education system at different points in the education process. The way National Assessment has chosen to do this is to monitor three in-school age groups--9-, 13- and 17-year-olds--to see if they are gaining or losing ground in comparison to predecessors of the same age groups in prior assessments. When resources permit, the Assessment also includes a sample of young adults aged 26 to 35 and a sample of 17-year-olds who are not enrolled in a secondary school.

To measure changes in performance, approximately one-half of the exercises assessed in a given subject area are reassessed under almost identical administrative conditions. Each respondent completes only one booklet of items (exercises) of about 45 minutes in length. From 20 to 40 booklets are administered in an assessment year. Since a national probability sample of between 2,000 and 2,600 students respond to each booklet, the total sample size can vary from 40,000 to 100,000 students.

In each assessment, 13-year-olds are assessed in October through December, 9-year-olds in January and February, and 17-year-olds in March and April. Thus, the amount of school experience in terms of time spent in school is approximately the same in each assessment for each age group. (Young adults and out-of-school 17-year-olds are typically assessed during the summer months.)

The exercises for each assessment are administered by a professional data collection staff to minimize the burden on participating schools and to maximize uniformity of assessment conditions. Instructions and items are recorded on a paced audio tape and played back to students to reduce the potential effect of reading difficulties and to insure that all students move through the booklets at the same speed.

Multiple-choice items are scored by an optical scanning machine; open-ended items are hand-scored by trained scorers using scoring guides that define categories of acceptable and unacceptable responses. These scoring guides are developed following field testing of the items and then revised and refined during receipt of initial assessment data.

National Assessment reports estimated percentages of acceptable responses for single items. When a report indicates that "85 percent of the 17-year-olds gave an acceptable response," it means that an estimated 85 percent of the 17-year-olds would have given an acceptable response if all the 17-year-olds in schools across the country had been assessed. In addition to reporting national results, National Assessment provides data on the performance of various population subgroups within the national population: sex, race, region of the country, size and type of community lived in, grade and level of parental education. National Assessment aggregates percentages of success on various sets of items to provide data on changes in performance between assessments and on the differential performance of population subgroups.

Since various components of the overall National Assessment model affect analysis procedures and sometimes even limit the range of possibilities, this paper begins with brief descriptions of key activities. The first sections generally describe the methods used to develop objectives and exercises, draw the assessment sample, prepare materials for the administration of an assessment, administer the booklets and score the items. The latter sections contain discussions about the NAEP analysis of change. Section 10 contains some answers to questions frequently asked of Assessment staff concerning the analysis of Assessment data. The appendixes contain detailed information on reporting categories, background questions, smoothing approaches, nonresponse adjustments and analysis of data for complex surveys.

SECTION 1
OBJECTIVES DEVELOPMENT

The primary goal of National Assessment is to report on the current education status of young Americans and to monitor any changes in achievement over time. For each learning area to be assessed, NAEP asks consultants to develop objectives that define the subject area. For most areas this involves content by process matrix. In addition, they are asked to create guidelines for exercise writers by specifying examples of the knowledge, skills and attitudes to be assessed at each age level.

Education in America is a collaborative enterprise involving a great many people with widely differing philosophies. Providing information about education nationwide would be considerably easier if there were consensus about the means and ends of American education, but the fact is that Americans have conflicting and sometimes contradictory values regarding the goals of education and the means for achieving them. To develop an assessment that is truly national in scope and takes into account the diversity of curricula, values and goals across the country, National Assessment employs a consensus process for developing objectives, involving representation of many different groups of people.

Several types of consultants help either to develop new objectives or to review and revise existing NAEP objectives developed for prior assessments. Educators and scholars, including college and university specialists, classroom teachers, curriculum supervisors and persons involved in teacher education, make sure that the objectives include concepts, skills and attitudes that the schools should be teaching and those that they presently are teaching. Concerned citizens, parents and other interested lay persons have to agree that the objectives are important for young people to achieve, are free of education jargon and are not biased or offensive to any groups. Consultants are representative of the different regions of the country, minority groups, various types of communities, age levels, education philosophies, and so on.

SECTION 2
DEVELOPMENT OF EXERCISES

Exercises are developed to provide information about achievement levels for objectives and subobjectives or cells of a content matrix. Each exercise is designed so that its results either can be used alone, as an indicator of performance on a specific task, or used in conjunction with results from other exercises to give a more general picture of achievement levels. In addition, except in cases of special interest, exercises are developed to provide information suitable for analyzing changes in performance over time. In this sense they must be replicable and a valid measure of achievement over time.

Most item writing is done by groups of people knowledgeable in the subject area. Individuals generate items, which are then reviewed and revised by other professionals in the subject area. Reviewers consider age-level appropriateness, accuracy of content, how well the item measures a question or objective, and readability. Exercises passing the group review are edited by the National Assessment staff to fit NAEP format and technical requirements.

The exercises produced by the writing groups are field tested in schools across the country to discover potential problems in wording, directions or administration procedures and to collect item statistics, timing information and scoring information. "Tryout" schools are selected to represent high- and low-income communities as well as more typical communities.

After the initial exercise pool is developed and field tested, consultants review the exercises and accompanying tryout data to insure that content areas are adequately covered. These people edit the existing exercises and generate new ones, which again, are field tested.

Finally, all items considered appropriate for inclusion in an assessment are reviewed in a series of conferences by numerous consultants. Exercises for each age group are reviewed by a variety of subject-matter specialists including classroom teachers. Lay citizens, representing a variety of occupations and interests, also review the exercises, checking for any type of bias and considering the general importance of each exercise. For more detailed

information, see The National Assessment Approach to Objectives and Exercise Development (1980).

SECTION 3
PREPARATION OF ASSESSMENT MATERIALS

Following the selection of exercises to be included in an assessment, National Assessment staff group and sequence items into exercise booklets. Since students at different ages receive somewhat different sets of exercises, booklets are constructed separately for each age level.

The following constraints are observed in preparing exercise booklets:

1. Each booklet contains exercises of varying difficulty so that students will not become bored by many easy exercises or discouraged by many difficult exercises.

2. Booklets are designed to be parallel, insofar as possible with respect to the number of different objectives measured. Exercises measuring a particular objective are scattered throughout the booklets so that many different students will respond to questions related to a particular objective.

3. Exercises cannot cue other exercises. In other words, the answer to one exercise cannot be contained in another exercise in the same booklet.

4. Each booklet is timed so that it will take no more than 45 minutes--the length of a typical class period--of a student's time. Booklets contain approximately 30 to 35 minutes of exercise time and an additional 10 to 15 minutes of introductory material, instructions and background questions.

National Assessment makes every effort to minimize difficulties connected with the testing situation so that results will be, as nearly as possible, an accurate reflection of what students know and can do. For example, students mark their answers directly in the assessment booklets, not on separate answer sheets. It is felt that this procedure will reduce the possibility of errors in marking answer sheets, especially for the younger students. To minimize guessing, students are encouraged to write "I don't know" on the answer line for open-ended questions or to select the "I don't know" response option included with each multiple-choice exercise if they feel they do not know the answer to a question.

Paced audio tapes are prepared for each exercise booklet to minimize the effect of any reading difficulties and to insure that all students move through the booklets at the same speed. In addition, the use of tapes helps to insure uniform assessment conditions across the country.

SECTION 4
SAMPLING

The target populations for each assessment consist of 9-, 13- and 17-year-olds enrolled in either public or private schools at the time of the assessment who are not functionally handicapped to the extent that they cannot participate in an assessment. Specific groups excluded are: non-English-speaking persons, those identified as nonreaders, persons physically or mentally unable to respond, and persons in institutions or attending schools established for the physically or mentally handicapped.

National Assessment does not follow up specific individuals from one assessment to the next. However, in each assessment year, participants are carefully selected to represent each age level. The definitions of the target populations are identical in each assessment. However, the sample design used to obtain representative samples of the target populations may be modified somewhat between assessments.

National Assessment uses a deeply stratified, three-stage, national probability sample design with oversampling of low-income and rural areas. In the first stage, the United States is divided into geographical units of counties or groups of contiguous counties meeting a minimum population size requirement. These units, called primary sampling units (PSUs), are stratified by region and size of community. From the list of PSUs, a sample of PSUs is drawn (without replacement) with probability proportional to populaton size measures, representing all regions and sizes of communities. Oversampling of low-income and extreme-rural areas is first performed at this stage by adjusting the estimated population size measures of those areas to increase sampling rates. Within PSUs, Census Employment Survey Data are used to further delineate and oversample low-income areas. Counties with high proportions of rural families are also oversampled.

In the second stage, all public and private schools within each PSU selected in the first stage are listed. Schools within each PSU are selected (without replacement) with probabilities proportional to the number of age-eligibles in the school.

The third stage of sampling occurs during the data collection. A list of all age-eligible students within each selected school is prepared. A simple random selection of eligible students (without replacement) is obtained, and exercise booklets are administered to selected students by specially trained personnel.

Each respondent in the sample does not have the same probability of selection, primarily because some subpopulations are oversampled and adjustments are made to compensate for some schools' refusal to participate and for student nonresponse. The selection probability for each individual is computed, and its reciprocal is used to weight each response in any statistical calculation to compensate for unequal rates of sampling and to insure proper representation in the population structure.

The number of PSUs, schools within PSUs and students within schools is determined by optimum sampling principles. That is, a sample design is selected that will achieve the maximum precision for a given level of resources. The current design uses about 75 PSUs in each assessment and 1,700 schools. See Appendix F for a discussion of the effects of the sample design upon the analysis.

SECTION 5
DATA COLLECTION

Participation in the National Assessment is voluntary. NAEP makes every effort to encourage the schools selected in the sample to participate in the assessment, and National Assessment and Research Triangle Institute staffs have obtained high rates of school cooperation (over 90 percent). Student cooperation rates are also high, especially for 9- and 13-year-olds (90 percent, 85 percent, respectively). Special follow-up procedures employed in recent years secure over 80 percent cooperation at age 17. For more information, see Access to School Districts, Schools and Nonstudents (1980). The effect of student nonresponse is discussed in Appendix E.

A professional data collection staff from the Research Triangle Institute, Raleigh, North Carolina, is used so that the burden on participating schools will be minimized and to assure uniform administration conditions.

National Assessment protects the anonymity of each respondent. Students' names are listed with their booklet identification number to enable verification if necessary. However, these lists do not leave the schools and are destroyed six months following the assessment.

School officials are asked to respond to a questionnaire asking about the enrollment in various grades, the types of communities in which the students live and the general occupational levels of the people in the community. The assessment administrator codes each student's birth date, sex, grade, racial/ethnic classification and identification number on his or her booklet. Six different racial classifications are used: white, black, Spanish heritage, American Indian or Alaskan native, Pacific Islander or Asian, and unclassified. Forms used to collect background information are found in Appendix B.

Each age group is assessed at approximately the same time of the school year in each assessment. As noted previously, 13-year-olds are assessed in October through December, 9-year-olds in January and February and 17-year-olds in March and April.

Following data collection, assessment administrators send completed booklets to the scoring contractor, Westinghouse DataScore Systems, Iowa City, Iowa. Booklets are counted and quality-checked to verify that correct administrative procedures were followed. Coded identification information is also checked for accuracy.

SECTION 6
SCORING

Scoring and conversion of the data into machine-readable form is contracted to Westinghouse DataScore Systems. Responses to multiple-choice exercises are read directly by optical scanning machines. A special staff scores responses to open-ended exercises by hand. Scorers use well defined guides to categorize responses and code the information into ovals that can be read by the optical scanning machine.

Scorers are carefully trained in the use of the scoring guides by scoring sample responses until they feel comfortable using the guides and categorizing the data reliably. To further ensure the quality and consistency of scoring open-ended exercises, quality-control checks are conducted at regular intervals. If discrepancies in scoring become apparent, scorers are retrained and, on some occasions, responses are rescored.

To measure changes in performance accurately, all responses to open-ended items collected in successive assessments of the same subject area either are categorized at the same time by the same scorers or calibration samples are used.

SECTION 7
DATA ANALYSIS

Measures of Achievement

National Assessment reports the performance of groups of students, not individuals. The basic measure of achievement reported by the Assessment is the percentage responding acceptably to a given item. This percentage is an estimate of the percentage of 9-, 13- or 17-year-olds who would respond acceptably to a given item if every 9-, 13- or 17-year-old in the country were assessed.

Percentages of acceptable responses are used because each item is designed as a separate measure of some aspect of an objective or subobjective. The purpose of National Assessment is to discover if more or fewer people are able to answer these items correctly.

In addition to providing national results, National Assessment reports on the achievement of various subpopulations of interest. Groups are defined by region of the country, sex, race, size and type of community lived in, grade and level of parents' education. Results for some additional variables are also analyzed. The definition of reporting groups is found in Appendix A. Forms used to obtain basic background information are presented in Appendix B. Subject-area-specific background questions are included in Appendix C. Appendix D discusses the smoothing of subpopulation proportions across booklets and years.

Procedures for estimating percentages of acceptable responses to exercises are dependent on the sample design. Each response by an individual is weighted and multiplied by an adjustment factor for nonresponse (see Appendix E). An estimate of the percentage of a particular age group that would have responded to an exercise acceptably if the entire age group were assessed is defined as the weighted number of acceptable responses divided by the weighted number of all the responses. A similar ratio of weights is used to estimate percentages of acceptable responses for reporting groups or subpopulations of interest.

The difference between the percentage of acceptable responses for a reporting group and that of the entire age group (nation) on an exercise describes the performance of any reporting group relative to the entire age group. This difference is a positive number if the group achieves a higher percentage than the entire age group and is a negative number if the group achieves a lower percentage. For example, a group performance of +1.8 indicates that the percentage of acceptable responses for the group is 1.8 percentage points higher than the national percentage of acceptable responses for a particular age level.

Increases or decreases in the percentage of acceptable responses between two assessments of the same subject area are estimated by finding the difference between percentages obtained from each assessment. A positive difference indicates an increase, and a negative difference indicates a decrease in the percentage of students who responded in a particular way from one assessment to the next. These differences, or change measures, are used to indicate trends in achievement, or performance, for an age level or subpopulation of interest. Changes between assessments in group differences from the national performance can be used to compare the trend of a group with the national trend.

To present a general picture of changes in achievement, National Assessment summarizes the performance for each assessment (either for the entire learning area or for some integral set of exercises) by using the mean, or arithmetic average, of percentages of acceptable responses to the exercises. During the first years of National Assessment, the median was used as the principal summary measure. However, the mean was chosen as the principal summary measure of change after extensive investigation showed empirically that it was more suitable for National Assessment change data than alternative measures. Twenty-two empirical distributions of change measures from the 1969-70 and 1972-73 science assessments were used to generate Monte Carlo simulations of sampling distributions for several measures of central location. In addition to the mean and median, other measures of central location that were considered in the simulation studies included the average of the extremes, two forms of biweighted estimates and three forms of weight-matching estimators described by John Tukey in the research paper "Some Considerations on Locators Apt for Some Squeezed-Tail (and Stretched-Tail) Parents" (1975). In almost every case, the sampling stability of the mean change was as good as or better than that of the other measures studied.

In addition, the mean is an easily understood and well-known statistic and has simple arithmetic properties useful for the analysis of differences or change measures--in particular, the difference between means is the same as the mean difference. This property allows us to accurately describe the mean change as the difference between mean percentages of acceptable responses from one assessment to the next. Mean percentages are used to simplify descriptions of change and descriptions of differences in group performances. Differences in mean performance between reporting groups and the nation are used to summarize the performance of various groups relative to the nation. NAEP does not intend that the mean percentage should be construed as an average test score.

Unless the items summarized in the mean percentages of acceptable responses are identical, the means of one age group should not be compared with the means of another, since their values reflect the difficulty of exercises in addition to the performance of the students. When only a few exercises are summarized by a mean, one should be especially cautious in interpreting results, since a small set of exercises might not adequately cover the wide range of potential behaviors included under a given objective or subobjective. The mean should be interpreted literally as the arithmetic average of the percentage of acceptable responses obtained from National Assessment samples on a specific set of exercises.

In considering National Assessment's achievement measures, it is the changes in performance between assessment years, among groups and among ages that are the most useful. By maintaining the same item or set of items in making these comparisons, we have a reasonable indicator of whether more or fewer people know or can do something judged important.

Estimating Variability in Achievement Measures

National Assessment uses a national probability sample at each age level to estimate the proportion of people who would successfully complete an exercise. The particular sample selected is one of a large number of all possible samples of the same size that could have been selected with the same sample design. Since an achievement measure computed from each of the possible samples would differ from one sample to another, the standard error of this statistic is used as a measure of the sampling variability among

achievement measures from all possible samples. A standard error, based on one particular sample, serves to estimate that sampling variability.

In the interest of sampling and cost efficiencies, National Assessment uses a complex, stratified, multistage probability sample design. Typically, complex designs do not provide for unbiased or simple computation of sampling errors. A reasonably good approximation of standard error estimates of acceptable response percentages is obtained by applying the jackknife procedure (Miller, 1964, pp. 1594-1705; Miller, 1968, pp. 567-582; Mosteller, 1968) to first-stage sampling units within strata. Standard errors for achievement measures such as group differences, mean percentages or mean group differences for a particular assessment year are estimated directly, taking advantage of features of the jackknife procedure that are generic to all of these statistics. Since samples for different assessments are independent, the standard errors of the differences in achievement measures between assessments can be estimated simply by the square root of the sum of squared standard errors from each of the assessments.

The standard error provides an estimate of sampling reliability for the achievement measures used by National Assessment. It is comprised of sampling error and other random error associated with the assessment of a specific item or set of items. Random error includes all possible nonsystematic error associated with administering specific exercises to specific students in specific situations. Random differences among scorers for open-ended items are also included in the standard errors.

Differences in performance between assessment years or between a reporting group and the nation are reported as statistically significant only if the differences are at least twice as large as their standard errors. Differences this large would occur by chance in fewer than 5 percent of all possible replications of the sampling and data collection procedures for any particular reporting group or national estimates.

Controlling Systematic Errors

Systematic errors can be introduced at any stage of an assessment--exercise development, preparation of exercise booklets, design or administration procedures, field administration, scoring or analysis. These nonsampling,

nonrandom errors rarely can be quantified, nor can the magnitude of the bias they introduce into our estimates be evaluated directly.

Systematic errors can be controlled in large part by employing uniform administration and scoring procedures and by requiring rigorous quality control in all phases of an assessment. If the systematic errors are the same from age to age or group to group, then the differences in percentages or mean percentages are measured with reduced bias because subtraction tends to cancel the effect of the systematic errors.

Similarly, the effect of systematic errors in different assessment years can be controlled by carefully replicating in the second assessment the procedures carried out in the first. Differences in achievement across assessment years will also be measured with reduced bias since subtraction will again tend to cancel systematic errors.

Although it is not possible for every condition or procedure to remain exactly the same between assessments conducted several years apart, National Assessment makes every effort to keep conditions as nearly the same as possible.

Reporting the Data

Each assessment generates a tremendous amount of data. To make these data as useful as possible to a variety of audiences, National Assessment generally provides data and reports its analysis results in the following forms.

Summary Reports

As noted, in addition to computing exercise-level data, National Assessment provides means for selected sets of exercises within a learning area assessment. These means are useful in comparing performance between assessment years and between reporting groups. Means are computed for a number of reporting variables, including region, sex, race/ethnicity, size and type of community, grade and level of parental education, as well as additional variables appropriate to the learning area. Besides publishing many of these results in tabular form, National Assessment prepares overview reports for the general public--including parents, classroom teachers, school administrators and

legislators--that summarize trends and highlight assessment results.

Although National Assessment does not interpret assessment results, it recognizes that data presented alone are often difficult to consider in perspective. Accordingly, NAEP frequently asks groups of educators to review and comment upon the results and includes their observations in the reports.

Public-Use Data Tapes

For those who wish to perform their own analyses of National Assessment data, NAEP makes data tapes of respondent-level data available for selected assessments. To protect the confidentiality of the respondents, all identifying information (school, district, state) has been deleted. The tapes also include documentation of exercises and classification schemes used by NAEP to summarize exercises. The tape is organized and documented to facilitate use with standard statistical packages.

Microfiche

National Assessment is in the process of making analysis results available on microfiche.

User Services

National Assessment provides some assistance to those wishing to use assessment items or to replicate assessment methodology. Those interested in receiving assistance should contact the National Assessment office.

SECTION 8
COMPUTATION OF MEASURES OF ACHIEVEMENT, CHANGES IN ACHIEVEMENT AND STANDARD ERRORS

Several measures of achievement that National Assessment uses in its reports are described in Section 7 of this document. The sample design is a complex, deeply stratified, multistage probability sample design. Measures of achievement are obtained by weighting individual responses appropriately. Reasonably good approximation of standard error estimates of these achievement measures can be obtained by applying the jackknife procedure to first-stage sampling units (replicates) within strata, using the method of successive differences and accumulating across strata.

In this section, the measures of achievement are first defined in algebraic form, followed by a description of the jackknife method used by National Assessment to estimate their standard errors.

Measures of Achievement

Based on the sample design, a weight is assigned to every individual who responds to an exercise administered in an assessment. The weight is the reciprocal of the probability of selecting a particular individual to take a particular exercise with adjustment for nonresponse. Since the probabilities of selection are based on an estimated number of people in the target age population, the weight for an individual estimates the number of similar people that individual represents in the age population. As explained in Appendix D, the weights are adjusted to reflect information from previous assessments on population distributions.

A sum of the weights for all individuals at an age level responding to an exercise is an estimate of the total number of people in that age population. A sum of weights for all individuals at an age responding correctly to an exercise is an estimate of the number of people who would be able to respond correctly in the age population if the entire population were assessed. These concepts also apply to any reporting group (e.g., defined by region, sex, and so

on) and category of response (e.g., correct, incorrect and "I don't know").

Let W_{ihk}^{e} = sum of weights for respondents to exercise e who are in reporting subgroup i and who are in the kth replicate of the hth sampling stratum, and

C_{ihk}^{ej} = sum of weights for respondents to exercise e who are in reporting subgroup i, who are in the kth replicate of the hth sampling stratum, and who selected response category j (e.g., correct foil) for the exercise.

Note that $W_{ihk}^{e} = \text{Sum}_{j} C_{ihk}^{ej}$

Then, summing k over the n_h sample replicates in the stratum h, and summing over the H sampling strata,

$$W_{i++}^{e} = \text{Sum}_{h=1}^{H} \text{Sum}_{k=1}^{n_h} W_{ihk}^{e}$$

estimates the number of eligibles in the population who are in subgroup i.

Similarly, $C_{i++}^{ej} = \text{Sum}_{h=1}^{H} \text{Sum}_{k=1}^{n_h} C_{ihk}^{ej}$

estimates the number of eligibles in the population who are in subgroup i and who would select response category j for exercise e.

An estimate of the proportion of the eligibles in the age population in group i who would select response category j on exercise e is:

$$(1) \quad P_{i}^{ej} = C_{i++}^{ej} / W_{i++}^{e}.$$

In the special case where the proportion of all age eligibles who would select response category j on exercise e is estimated, the index A (for ALL) will be used in place of i as follows:

$$(2)\quad P^{ej}_{A} = C^{ej}_{A++} / W^{e}_{A++}.$$

In National Assessment reports, the proportion in (1) multiplied by 100 is called the group percentage, and the proportion in (2) multiplied by 100 is called the national percentage. The difference between the proportion in subgroup i who would select category j on exercise e and the proportion in the nation is denoted by:

$$(3)\quad dP^{ej}_{i} = P^{ej}_{i} - P^{ej}_{A}.$$

National Assessment also reports the arithmetic mean of the percentage of correct responses over sets of exercises corresponding to the measures in (1), (2) and (3). These means are taken over the set of all exercises or a subset of exercises classified by a reporting topic or content objective. The mean percentage of correct responses taken over m exercises in some set of exercises corresponding to measures (1), (2) and (3) are, respectively:

$$(4)\quad \bar{P}_{i} = \frac{1}{m} \text{Sum}_{e}\, C^{e}_{i++} / W^{e}_{i++},$$

$$(5)\quad \bar{P}_{A} = \frac{1}{m} \text{Sum}_{e}\, C^{e}_{A++} / W^{e}_{i++} \quad \text{and}$$

$$(6)\quad \bar{dP}_{i} = \bar{P}_{i} - \bar{P}_{A}.$$

Note that the response category subscript j has been suppressed since the means are understood to be taken over the correct response category for each exercise.

Each of these six achievement measures is computed and routinely used in reports describing achievement data for any assessment. The simple difference in these measures between two assessments of the same exercise (or sets of exercises) provides six measures of change in achievement that are routinely used in National Assessment's change reports. The next section describes how standard errors are estimated for the 12 statistics routinely used in NAEP reports.

Computation of Standard Errors

In order to obtain an approximate measure of the sampling variability in the statistics (1) through (6), a jackknife replication procedure for estimating the sampling variance of nonlinear statistics from complex, multistage samples was tailored to National Assessment's sample design. Miller (1968, 1974) and Mosteller and Tukey (1977) provide information about the jackknife technique, while Folsom (1977) describes how the procedure is used in estimating standard errors for National Assessment's sample designs.

To demonstrate the computational aspects of this technique, consider estimating the variance of the statistic in (1)--the proportion of age-eligibles in subgroup i who would select response category j on exercise e.

This statistic is based on the data from all the n_h replicates in the H strata. Let $p_{i\text{-}hk}^{ej}$ be defined as a replication estimate of p_i^{ej} and constructed from all the replicates excluding the data from replicate k in stratum h. These replication estimates are computed as if the excluded replicate had not responded and a reasonable nonresponse adjustment is used to replace the data in replicate hk in estimating p_i^{ej}. Several choices for replacing the data in replicate hk are available. In order to obtain a convenient and computationally efficient algorithm for approximating standard errors, National Assessment replaces C_{ihk}^{ej} and W_{ihk}^{e} from the hkth replicate with corresponding sums from another paired replicate in the same stratum. The replicate

estimate is then computed. The replicate estimates to be used in the calculations are determined by arranging all the replicates in each stratum into successive pairs. That is, replicate 1 is paired with replicate 2, replicate 2 with replicate 3, 3 with 4, ... (n -1) with n and replicate n with replicate 1. n_h n_h

The contribution to the variance of p_i^{ej} by each pair of replicates is the change in the value of the statistic incurred by replacing the data from each replicate in the pair with the data from the other replicate in the pair and recomputing p_i^{ej} in the usual way. This produces two replicate estimates. Squaring the difference between these replicate estimates and then dividing by eight measures the contribution of this pair of replicates to the total variance. The sum of these contributions over all n_h successive pairs in the stratum is the contribution by stratum h to the total variance. The square root of the sum of the H stratum contributions is the estimate of the standard error of p_i^{ej}.

Algebraically, the two replicate estimates for the pair k, k+1 (where k=1,...n_h and n_h+1=1) are:

$$(7)\quad p_{i\text{-}hk}^{ej} = \frac{C_{i++}^{ej} - C_{ihk}^{ej} + C_{ih(k+1)}^{ej}}{W_{i++}^{ej} - W_{ihk}^{ej} + W_{ih(k+1)}^{ej}}$$

and

$$(8)\quad p_{i\text{-}h(k+1)}^{ej} = \frac{C_{i++}^{ej} - C_{ih(k+1)}^{ej} + C_{ihk}^{ej}}{W_{i++}^{ej} - W_{ih(k+1)}^{ej} + W_{ihk}^{ej}}$$

The contribution to the total variance from stratum h is:

$$(9) \quad \text{var}\left(P_{ih}^{ej}\right) = \frac{1}{8}\sum_{k}^{n_h}\left[P_{i-hk}^{ej} - P_{i-h(k+1)}^{ej}\right]^2$$

And, finally, an estimate of the standard error of p_i^{ej} is:

$$(10) \quad SE\left(P_i^{ej}\right) = \left(\sum_{h}^{H} \text{var}\, P_{ih}^{ej}\right)^{1/2}$$

Multiplying p_i^{ej} by 100 yields the percentage of response to category j. Multiplying $SE(p_i^{ej})$ by 100 yields the corresponding estimated standard error of the percentage.

In general, the jackknifed standard errors of the proportion estimates will be larger than the simple random sampling formula $(pq/n)^{1/2}$, where $p=p_i^{ej}$, q=1-p and n is the number of sampled respondents in subgroup i who took the exercise. The larger size of $SE(p_i^{ej})$ reflects mainly the loss of precision due to cluster-sampling of schools and students.

The standard errors for the achievement measures (2) through (6) are computed through a series of steps analogous to those followed in computing $SE(p_i^{ej})$. The most complicated step in computing standard errors occurs in forming the paired replicate estimates analogous to (7) and (8) for each successive pair of replicates. Once this bookkeeping chore is done, the computations (9) and (10) follow in a straightforward manner.

The standard errors for the differences between two assessments for any of the achievement measures (1) through (6) are computed as the square root of the sum of the squared standard errors from each of the separate assessments.

The size of the standard errors depends largely not only on the number of replicates and schools included in the sample, but also on the number of respondents in each of the reporting groups. Table 8-1 shows the average proportion of students responding to an exercise booklet for each of the reporting groups for each age. The proportions in Table 8-1 are for the 1977-78 assessment, averaged over booklets and smoothed over several adjacent assessments.

The size of the standard errors of the means of the achievement measures for sets of exercises is also influenced by the number of exercises in the exercise set and the number of booklets over which the items in the set are spread.

Table 8-1. Average Number of Respondents in Reporting Groups Taking an Item Booklet, By Age, for the 1977-78 Assessment

	Age 9	Age 13	Age 17
Nation	2,456	2,424	2,230
Region			
Northeast	552	547	516
Southeast	591	546	509
Central	764	770	731
West	549	561	474
Sex			
Male	1,234	1,210	1,085
Female	1,221	1,214	1,145
Race			
White	1,848	1,862	1,796
Black	433	396	306
Parental education			
Not graduated high school	204	302	334
Graduated high school	579	797	763
Post high school	772	942	1,020
Type of community			
Extreme rural	240	242	221
Disadvantaged urban	245	241	214
Advantaged urban	247	243	227
Size of community			
Big cities	640	685	612
Fringes around big cities	524	519	483
Medium cities	170	170	126
Small places	1,122	1,049	1,008
Grade			
3, 7, 10	634	652	303
4, 8, 11	1,755	1,704	1,675
12	--	--	220

SECTION 9
SOME ANALYSIS ALTERNATIVES FOR DETERMINING CHANGE IN STUDENT PERFORMANCE

NAEP estimates achievement based on the percentages of young people acceptably answering a given exercise or set of exercises; changes in achievement are the differences between the percentages of young people acceptably answering a given exercise or set of exercises from one point in time to another. This section presents brief discussions of alternatives that have been suggested for measuring change.

Longitudinal Versus Cross-Sectional Studies

As stated in the Introduction, the purpose of National Assessment is to measure the effect of the education system on students at different points in time. To do this NAEP uses a cross-sectional approach, in essence, taking a "snapshot" of achievement levels at various stages in the education process (ages 9, 13 and 17) at approximately five-year intervals. The pictures or results are then compared to see if there have been any changes.

It is not National Assessment's purpose to monitor individuals' performance or study how individuals progress through the education system. To do this would require a longitudinal approach with follow-up of the same individuals over time. The longitudinal approach with its concern with the lives and developmental patterns of individuals over a period of time facilitates the understanding of the effects of education programs, policies and practices. This kind of approach permits researchers to address some questions that the current cross-sectional approach of National Assessment cannot address.

Similarly, a single longitudinal study, which follows a sample over time, cannot address questions concerning changes over time in the education system as can a series of cross-sectional measures. Because a single longitudinal study of say four follow-ups could be more expensive than four cross-sectional studies, and because the cross-sectional approach is suited to addressing National Assessment's primary purpose, National Assessment has chosen to use the cross-sectional approach.

In addition, there are limitations associated with the longitudinal approach. One problem is the loss of representativeness of the sample in subsequent follow-up. This can happen in at least two ways. First, if one begins with a nationally representative sample of 9-year-olds and follows them up every four years for twelve to twenty years, then at each measurement point we will find that some of our originally sampled individuals will disappear from the sample for a variety of reasons and we are unable to locate them. Another way in which representativeness can be lost is that the persons included in the longitudinal sample are aware of the fact that they are part of a special study and in some ways may no longer be representative because of this fact.

An approach that would still permit National Assessment to monitor changes in the education system and also to address questions well suited to the longitudinal approach would be one of performing multiple longitudinal studies. This might be accomplished by beginning a new longitudinal study in a subject matter, say every four or five years, and using the first points of each longitudinal study just as we presently use the cross-sectional measures. This procedure, although probably difficult to implement, would greatly enhance the analytical possibilities for NAEP data. However, to embark on such an endeavor would require a rather dramatic increase in the resources devoted to National Assessment. A less ambitious effort would be to build into the current design a small longitudinal component.

Alternative Exercise Development and Release Strategies

There has been universal agreement that to measure change between two points in time, the same exercises should be readministered in the same way. Simple differences then constitute the change measure. To implement this analytical procedure, it was planned to retain and keep secure approximately 50 percent of the items from each assessment for the purpose of measuring change and make the remaining items available to the public. With each successive assessment, the released items would be replaced with new items. After the second assessment the items used to measure change would be released, while those administered for the first time in the second assessment would be held to measure any changes in the third assessment and so on through successive assessments.

This procedure is well received because updating the exercise pool for each assessment allows the exercises to remain consonant with current education goals. Also, states and districts are interested in using the released exercises to compare their student performance with that of the nation or a particular subpopulation. However, there is the question of how to start establishing trends across time from only successive pairs of links based on different exercises.

One point of view is simply to consider the most recent change measure as the important one, because it should address the domains currently emphasized in schools.

An approach to exercise development that alleviates this analytical concern is that of sampling from large exercise pools. Initially, an extremely large pool of thousands of exercises is developed that is so representative of the subject area as to almost be considered a definition of the subject area. Thus, it could be made available to the public. Items would not be readministered from assessment to assessment, but for each successive assessment a representative sample of items would be drawn and administered. Although this procedure does enable the analysis of trends, there is still the state of the art of exercise development to be considered. It would be difficult to accept these different sets of items as equally valid and equally difficult measures of achievement. In addition, to satisfy subject-area consultants, the pool would have to be revieved and updated almost as often as NAEP presently develops new exercises. Finally, this approach would involve extreme expenditures to develop the initial pools.

Readministering the same items at multiple points across time also enables the computation of trends. However, for any measurement of change, care must be taken not to include time-bound exercises or exercises no longer deemed important (or taught) to avoid spurious trends. It is unlikely that any given set of exercises would ever endure for long periods of time as a valid measure of achievement for a subject area.

In actuality, National Assessment has retained closer to two-thirds of the items included in some assessments for the purpose of measuring change. Thus, some exercises have been administered at three points in time and some will be readministered a fourth time.

So far, the analysis approach adopted by National Assessment generally has been to report change across three points in time using all the pertinent information available to try to establish trends.

However, given finite resources such continuous readministration of exercises is limiting flexibility, both for allowing refurbishing of the pool and providing released materials for those interested in using NAEP materials for their own purposes. Also, change across the most recent two points in time may still be of primary interest, since it is based on newer exercises that may be more comprehensive and pertinent in terms of current education programs and curriculum. It is planned that as the number of points in time increases, curve-fitting approaches will be employed to establish trends across successive links in time.

Latent Trait Models

Although latent trait models have actually been around for some time, it is only within the last five to ten years that interest has been growing in the application of this approach--primarily for testing individuals. In practical applications, the research usually specifies the mathematical form of its characteristic curves and obtains estimates of the item parameters that describe the curves. Item characteristic curves represent the probability of a correct answer to an item expressed as a function of ability.

The use of latent trait models has been suggested as an approach for the measurement of change in group performance. For simple changes between two points in time, this approach requires many more assumptions and calibration efforts than does the approach currently used. It is unlikely that the assumptions required for any latent trait model will ever be completely satisfied by data sets that meet the needs of National Assessment. Just how robust latent trait models are is a question that has not been thoroughly explored in the National Assessment context.

For long-term trends and cross-area analysis of change, this approach may offer some advantages. National Assessment is currently exploring possible ways of utilizing latent trait methodology.

Other Approaches

If the basic assessment design were changed, a wide variety of other approaches to measuring change could be considered. National Assessment is currently being reviewed by a year-long study funded by a consortium of foundations--Carnegie, Ford, Kettering, Spencer and Mott. This effort, known as the Assessment of the Assessment, will make recommendations for future changes that will be considered by the Assessment Policy Committee (see Question 1, Section 10).

SECTION 10
QUESTIONS FREQUENTLY RAISED CONCERNING NAEP ANALYSIS

1. How committed is NAEP to the current model? What kind of realistic changes can be made without disrupting the "change" information? Is NAEP willing or able to change the model if it can be shown that more usable data would result?

The current model was developed with the aid of a panel of statistical and measurement experts (see Appendix G for a list of the members of the Analysis Advisory Committee). It has been periodically reviewed over the years and is currently undergoing an extensive review funded by a consortium of foundations (see Section 9 for a discussion of the review). Recommendations from this review will be presented to the Assessment Policy Committee. It is true that major changes in the model could disrupt the continuity of change measures. However, changes such as abandoning exercise level data and concentrating on summaries could be accommodated easily.

2. Why doesn't NAEP analyze its data to provide information on cause and effect? Why not develop hypotheses to be tested with the data?

NAEP is basically a sample survey of individuals' responses to various questions. While surveys frequently provide data that can be obtained in no other way, they do present certain difficulties in analysis. Since the data are from a survey, are observational in nature, and do not result from the control of variables (as do data from designed experiments), the results can lead to misinterpretations. In experimental research it is often possible to infer causation because the researcher has controlled many individual variables and used treatment randomization to eliminate any systematic effects of uncontrolled variables. In surveys, because the values of the various variables are not controlled and because the variables tend to be related to each other and to other extraneous variables, such control is not generally possible. Because correlation, in itself, does not prove causation, the direction of causation is often unclear for many variables from survey data.

In applying standard statistical tests, it is often the case that undesirable confounding occurs between pertinent variables and other, perhaps extraneous, variables; hence, tests of hypotheses performed in the presence of these, often unremovable, confounding variables require very careful interpretation. This is not to say that using tests of hypotheses is not useful, since it can provide valuable insights into potential mechanisms that are operating. Further experimental research may then clarify the actual causative agents. NAEP makes available public-use data tapes so that secondary researchers can explore the data in this way. (For more information about the availability of public use data tapes, see Question 14.)

To summarize, NAEP data should be examined in an exploratory, rather than confirmatory, framework. (For additional discussions of this topic, see Tukey, 1977; Appendix F; and Contributions of the National Assessment to Understanding the Problems of Literacy and Equity, 1980.)

3. Why doesn't NAEP use a longitudinal sample to monitor change rather than separate cross sectional samples of students?

   See Section 9 for a discussion of this topic.

4. The NAEP reporting of average performance by race characterizes all black students as performing less well than whites when in fact many black students perform as well as the top white students. Could the analysis be enhanced to present a clearer picture of achievement by race?

In the future NAEP will report six categories of achievement by race in addition to reporting average performance for the group (see Appendix A and Question 15 for more information concerning the achievement class variable).

5. NAEP reports data for various disadvantaged subgroups--minorities, disadvantaged urban (low metro, inner city), extreme rural and students whose parents have not graduated from high school. Yet, for example, when regional data are presented, no allowance is made for the fact that the Southeast has a higher proportion of disadvantaged students than other regions. How can you evaluate the effectiveness of the education process for groups with widely differing compositions?

The problems facing the schools in the Southeast are best shown by the actual data. Data adjustments should be used to better indicate the effectiveness of the education process. Unfortunately, no ideal adjustment has emerged for this task. This is primarily because of unavailability of all the characteristics to adjust, lack of knowledge of what all of these characteristics are and how they interact, and less than perfect measurement of the ones that are available. NAEP has presented adjusted data by using Tukey's "balancing" procedure (Science: Results by Race, Parental Education, Size and Type of Community, 1973). These data showed that the Southeast deficit was reduced when other factors were taken into account.

6. Why doesn't NAEP make use of latent trait models in the analysis?

   See Section 9 for a discussion of this topic.

7. Why does NAEP only assess age groups rather than both grade and age or grade only?

   Both grade and age data would be useful and would be collected if resources were available to do so. Age was chosen primarily because it is difficult to anticipate grading conventions for the future and also because it can be measured precisely.

8. The NAEP sample seems to be extremely complex. Why not use a simple random sample of students?

   Complex samples arise purely for economic reasons. For a given level of resources, the NAEP sample is many more times efficient in terms of providing information than an affordable simple random sample. By concentrating data collection efforts in the primary sampling units, the NAEP sample is several times larger than a simple random sample of the same total cost.

9. Could NAEP provide data for other regional breakdowns such as for the Great Plains, Appalachia, etc.?

   Yes. NAEP can provide data for a variety of other regional breakdowns. This would entail special tabulations with the cost to be borne by the requestor.

10. Could NAEP provide student-, school-, district- or state-level data?

The NAEP sample is in general not large enough to be representative at state and district levels. A much more extensive sample would be required to provide this data. The Assessment Policy Committee has decided not to release identifying data for the students, schools, districts or states assessed because of confidentiality commitments.

11. Why doesn't NAEP interpret its results or provide more contextual data in reports?

The Policy Committee of NAEP decided at the time of the 1969-70 assessment that National Assessment would not interpret its results. They made this decision because National Assessment would be in a position to manipulate school curriculum by developing, administering and interpreting an assessment. In addition, National Assessment generally has not had the subject-area expertise on staff that would be required to make an interpretative analysis of a subject-area assessment.

Instead of making interpretations of assessments, NAEP has encouraged readers to make their own interpretations and has encouraged teams of qualified experts to make published interpretations. National Assessment tries to be sure that all necessary technical measurement, sampling and administration information is made available to those making interpretations. Following the 1969-70 assessment, qualified experts were commissioned to write interpretations, which were published separately from National Assessment reports. In more recent years, however, it has been a fairly common practice for NAEP to convene a panel of qualified experts whose interpretations appear as separate chapters of the reports National Assessment publishes. Usually these panels of experts represent, at least in part, the associations of professional educators for the content area in question.

12. Data concerning background questions appear somewhat limited. Could NAEP provide additional information from existing data? Could additional background questions be included in future assessments?

As explained earlier, National Assessment routinely analyzes and reports data for a number of subpopulations--region, race, sex, size and type of community, level of parental education and grade. NAEP has generated data for some two-way cell analyses, but does not have a sufficient sample size for three-way analyses. The information necessary to compute these results is collected by the field staff or from school officials (see Appendix B).

In addition, NAEP has asked and plans to continue asking respondents a variety of background questions, some of which provide data applicable to all areas and some only appropriate to specific subject areas. Therefore, particularly for 17-year-olds attending school, NAEP could provide much additional information from existing files. For example, some data on more diverse groups such as bilingual students, Asians and American Indians are available. NAEP could also provide information concerning reading material in the home, amount of television watched and a variety of instructional practices in the school (see Appendices B and C). In NAEP's present financial situation, many analyses must take a somewhat secondary position. National Assessment hopes for additional resources that will enable additional analysis of background information and also that external researchers will take advantage of the public availability of NAEP data files to pursue additional analyses of this important information (see Question 14).

National Assessment would like to collect more background information, for example, about students' future education plans, the various influences on education attitudes, and more curriculum and program information from both students and teachers. The reason why NAEP does not include more background questions in each assessment is partly philosophic but primarily economic. For any given assessment, resourses limit the amount of information that can be gathered. There is always the question of trade-offs between particular background questions and collecting enough achivement data to provide adequate coverage of subject-matter domains. In order to collect more information, NAEP would either have to increase administration times, thereby increasing the overall burden on respondents, or increase sample size. Also, the

reliability of collecting some information through paper and pencil techniques must be considered. For example, one might want to find out exactly which textbooks students have used or find out students' IQs. However, it is unlikely that the information supplied by the student would be very accurate. Such information would have to be collected using different and sometimes extremely expensive techniques.

In considering the viability of asking additional questions, other issues must also be considered. Depending on the nature of the information desired, there are problems with the privacy of respondents. Some questions that researchers would like to ask are prohibited by privacy legislation, while other questions of interest are considered sensitive in certain areas of the country. There have been instances where particular questions currently asked by NAEP have been deleted by specific states or school districts. Finally, it must be remembered that NAEP surveys national samples of students and some questions could be more efficiently asked and analyzed in a separate research effort (see Question 2).

13. Why doesn't NAEP set performance standards for subject areas?

NAEP understands that setting performance standards would enhance the utility of the data for some audiences. This would enable NAEP to provide literacy rates that could be used by policy makers, comparative information on subject areas that could be used by curriculum decision makers to know which areas need the most attention, guidelines for proficiency levels that could be used by states and local districts, and so on. However, such a procedure is technically difficult, politically volatile and fairly expensive. As many states and local districts have found, setting defensible criteria for required performance levels is a very difficult task. To make such statements for the nation and have them be both equitable and meaningful could well be impossible (see Glass, 1977; Shepard, 1976.

The setting of absolute standards is often circumvented in education, with the more common practice being some type of relative approach where students are ranked. For example, a particular student is designated as in the top ten or at the 85th percentile. National Assessment currently uses the approach of comparing group achievement levels attained in a particular assessment to those levels attained on the same questions in earlier assessments. Thus, the previous level is used as the standard.

Prior to National Assessment even considering some standard setting procedure, the question of the meaningfulness and utility of standards at a national level would have to be carefully considered by the education community. It is difficult enough for a teacher to stipulate that every student, without exception, should know a specific piece of information. It is even more difficult to make those types of decisions consistently across subject areas. One is immediately faced with a type of interjudge reliability issue for which there are no established accurate measures. Are reading educators requiring an equitable level of performance as compared with mathematics educators? Furthermore, should these standards be "equal" or is reading in some sense "more important" than mathematics? In addition to the complex problems of expert agreement within a subject area and of comparable experts or judges to set standards across different subject areas, National Assessment also faces the problem of having different measures for different subjects at different points in time.

Even though no practical technology exists for setting performance standards that would be defensible to all NAEP constituents, there is still a demand for such data. Some procedure could be used that incorporates existing theories as well as experts' experiences and judgments. The necessity then remains to weigh the utility for decision makers of controversial standards against the many possible uses of NAEP's limited resources. In some cases, NAEP has decided to provide information that could be considered implicitly related to setting performance standards. For example, NAEP provides information showing performance levels or changes in performance levels for different objectives or topics within subject areas for specific age groups. Also, for all open-ended responses decisions are made about what is an acceptable level of performance. The primary trait scale, used for scoring certain writing tasks, suggests that a "1" is inadequate, a "2" is minimally adequate, a "3" is competent and a "4" is excellent. Finally, for many subject areas, consultants have established criteria related to attitudes.

14. How and in what form are NAEP data available for secondary researchers?

NAEP data are available to secondary researchers in several forms. Substantial quantities of summary data are provided in printed reports and microfiche. Most of these materials may be obtained through the Government Printing Office, ERIC or National Assessment. A much larger quantity of unpublished data is archived at the National Assessment offices in Denver. These data are available for use on-site and are reproduced on demand on a cost-recovery basis.

Respondent-level data files are also available for use by secondary researchers. Tape files and documentation for most assessments prior to 1976-77 are available, but in a form that most external users have found difficult to use.

With NIE and National Science Foundation support, NAEP has begun a major effort to make the data more accessible to secondary users. Problems with existing NAEP user tapes were studied, existing archiving systems and standards were investigated and a new public-use data tape system was designed. Implementation problems included balancing full access to data versus test security and confidentiality pledges, as well as trade-offs between production and user costs. Major features of the new system include: machine-readable documentation written specifically for secondary users, codebooks written in familiar formats, all numeric data files with uniform missing data and nonresponse codes, complete SPSS data definition cards, compatibility with other major statistical analysis packages, minimum user start-up costs, and compatibility with both university and Federal archiving standards.

By the end of September 1980, 88 files from the 1975-76 to 1977-78 assessments will have been produced and documented. A subset of files from the 1972-73 mathematics assessment, appropriate for trend analyses from 1972-73 to 1977-78 will also have been generated. The number of files, by assessment year, learning area and age level follow.

Number of Files Contained on
National Assessment Public-Use Data Tapes by
Assessment Year, Subject Area and Age Level

Year	Subject Area	Number of Files Age 9	Age 13	Age 17	Age 26-35
1977-78	Mathematics	7	11	12	
1976-77	Science	7	10	11	
	Energy, Health, Reading and Science				4
1975-76	Citizenship/Social Studies	5	8	9	
	Selected Supplemental Mathematics		2	2	
1972-73	Mathematics Change Exercises	7	9	11	

15. Are there any new analyses NAEP will be implementing in the near future?

The reporting of data by achievement class is a new enhancement to the analysis of change that will be used in the upcoming reading change report (see Appendix A for a description). Basically the achievement class variable partitions the sample of respondents by knowledge levels.

NAEP will primarily be using the achievement class as follows:

a. Measurement of Change. Does change in performance over time occur uniformly or differentially within each achievement class? Given a 3 percent change in performance from Year 06 to Year 10 (for a group or the nation), is there also about a 3 percent change in the performance for each of the achievement classes or is there a greater change (for example) in the lower achievement classes?

b. Exploration of the Characteristics of Low Achievers. Our current approach focuses on the achievement levels of various reporting groups but does not answer the question of what is the composition of low and high achievers in terms of various demographic characterisitcs. The achievement class variable will help to answer this question.

16. How may NAEP data analysis be enriched through comparisons with data from other local, state or national data bases?

The National Assessment data could be coded by school, district, county, zip code and state. Other data bases such as census data, ELSEGIS, IRS tapes, county records, and so on provide a variety of information at these levels. Correlational analysis of achievement with these other variables would perhaps provide valuable insights into potential contributors to improved education. Because of confidentiality considerations, analyses involving these levels would have to be conducted under special arrangements with National Assessment.

17. Couldn't changes NAEP observes be simply the result of "regression to the mean?"

Regression to the mean is a phenomenon that can occur when sample selection is based on prior performance. Since NAEP uses independent cross-sectional samples, the data are not subject to this problem.

18. Why doesn't NAEP use existing tests for some booklets or construct the booklets to be more like tests with items measuring the same concepts clustered together?

Subject-matter specialists have generally agreed that existing tests do not meet National Assessment standards for breadth and depth of coverage. Also, NAEP constructs each item to provide useful information if analyzed separately. Thus, NAEP item development is conducted from a slightly different perspective than it would be if the goal were to develop tests (see The National Assessment Approach to Objectives and Exercise Development, 1980).

Booklets containing uniform subject matter would enhance research possibilities, but detract substantially from NAEP's precision of change estimates. By distributing items measuring the same concept across a number of booklets, sample sizes used for estimates of mean performance are increased by 2,500 with each additional booklet. Also, respondents find a variety of items more interesting and considering the voluntary nature of the assessment participation rates must be taken into account.

SECTION 11
CONCLUSIONS AND RECOMMENDATIONS

This paper discusses in detail the approaches implemented in the analysis of National Assessment data. The purpose of this effort was to review these approaches, study alternatives and make recommendations for improving these procedures.

In preparing this paper no dramatic changes emerged as being obvious improvements. However, this paper will be available as background information for the extensive review of the assessment being sponsored by a consortium of foundations that is discussed in Section 9.

APPENDIX A

DEFINITIONS OF NATIONAL ASSESSMENT REPORTING GROUPS

In addition to reporting change results for all 9-, 13- and 17-year-old students in the United States, National Assessment reports results for a number of population subgroups.

Definitions of the subgroups follow.

Region

The country has been divided into four regions: Northeast, Southeast, Central and West.

Sex

Results are reported for males and females.

Race

Results are presented for blacks, whites and Hispanos. (Data for Hispanos are reported only for sets of exercises, not for individual items, because of small sample sizes.)

Level of Parental Education

Three categories of parental-education levels are defined by National Assessment, based on students' reports. These categories are: (1) those whose parents did not graduate from high school, (2) those who have at least one parent who graduated from high school, and (3) those who have at least one parent who has had some post-high-school education.

Type of Community

Communities in this category are defined by an occupational profile of the area served by a school as well as by the size of the community in which the school is located.

Advantaged-urban communities. Students in this group attend schools in or around cities having a population greater than 200,000 where a high proportion of the residents are in professional or managerial positions.

Disadvantaged-urban communities. Students in this group attend schools in or around cities having a population greater than 200,000 where a relatively high proportion of the residents are on welfare or are not regularly employed.

Extreme-rural communities. Students in this group attend schools in areas with a population under 10,000 where many of the residents are farmers or farm workers.

Size of Community

Big cities. Students in this group attend schools within the city limits of cities having a 1970 census population over 200,000.

Fringes around big cities. Students in this group attend schools within metropolitan areas (1970 U.S. Bureau of the Census urbanized areas) served by cities having a population greater than 200,000 but outside the city limits.

Medium cities. Students in this group attend schools in cities having a population between 25,000 and 200,000, not classified in the fringes-around-big-cities category.

Small places. Students in this group attend schools in communities having a population less than 25,000, not classified in the fringes-around-big-cities category.

Grade in School

Results are categorized for 9-year-olds in the third or fourth grade; 13-year-olds in the seventh or eighth grade; and 17-year-olds in the tenth, eleventh or twelfth grade.

Modal Grade by Region

Results are categorized for 9-, 13- and 17-year-old respondents in grades four, eight and eleven, respectively, who live in the Northeastern, Southeastern, Central or Western regions of the country.

Modal Grade by Community Size

Results are categorized for 9-, 13- and 17-year-old respondents in grades four, eight and eleven, respectively, who live in big cities, fringes around big cities, medium cities and small places.

Modal Grade by Sex

Results are categorized for 9-, 13- and 17-year-old males and females in grades four, eight and eleven, respectively.

Achievement Class

The achievement class variable classifies each respondent for a given age and year into one of a number of categories (achievement classes) based on the respondent's estimated standing in the population in terms of achievement. The classes form a partitioning of the population by achievement and are defined as follows:

Class 1: The lowest eighth (students in the lowest 12.5 percent of the population on achievement).

Class 2: The next-to-lowest eighth (students with higher achievement than 12.5 percent and lower achievement than 75 percent of the population).

Class 3: The next-to-lowest quarter (achieve higher than 25 percent and lower than 50 percent of the population).

Class 4: The next-to-highest quarter (achieve higher than 50 percent and lower than 25 percent of the population).

Class 5: The next-to-highest eighth (achieve higher than 75 percent and lower than 12.5 percent of the population).

Class 6: The highest eighth (students in the upper 12.5 percent of the population on achievement).

For our purposes, the measure of achievement for an individual will (generally) be the person's package score (mean percent correct) over all items from a particular subject area, such as reading, science, mathematics, and so on. The package score for an individual is the ratio of the number of correct responses over the number of items (including items where the "I don't know" foil was selected).

APPENDIX B

FORMS USED TO OBTAIN BACKGROUND INFORMATION

This appendix includes the basic forms to be used in the next assessment to collect background information from school officials and respondents. Following are a listing of the forms included.

School Principal's Questionnaire

Standard Background Information Form for 9-Year-Olds

Standard Background Information Form for 13-Year-Olds

Standard Background Information Form for 17-Year-Olds

School Principal's Questionnaire

This report is authorized by law (20 U.S.C. 1221 c-1). While you are not required to respond, your cooperation is needed to make the results of this survey comprehensive, accurate, and timely.

Primary Sampling Unit ☐☐ School Number ☐☐☐☐☐☐☐☐☐

Age Group(s) 9 13 17

PLEASE PRINT

Name of School _____

Address of School _____
(Street)

_____
(City) (State) (Zip Code)

Name of School Principal _____

Name and title of person completing the form if other than school principal

Name _____ Title _____

1. What is your best estimate of the current enrollment and the average daily attendance by grade of your school (current school year)? (Enter zeros for grades not served by your school.)

Grade	K	1	2	3	4	5	6	7	8	9	10	11	12
Enrollment													
Average Daily Attendance													

2. Approximately what percentage of the students attending your school live in each of the following areas?

_____% A In a rural area (less than 2,500)

_____% B In a town of 2,500 to 10,000

_____% C In a town of 10,000 or more

_____
100% (Items A-C should add to 100%)

3. Approximately what percentage of the students attending your school are children of

____% A Profcssional or managerial personnel

____% B Sales, clerical, technical or skilled workers

____% C Factory or other blue collar workers

____% D Farm workers

____% E Persons not regularly employed

____% F Persons on welfare

____ (Items A-F should add to 100%)
100%

4. Approximately what percentage of the students attending your school are

____% A American Indian or Alaskan Native

____% B Asian or Pacific Islander

____% C Hispanic, regardless of race (Mexican, Puerto Rican, Cuban, Central or South American or other Spanish culture or origin)

____% D Black and not Hispanic

____% E White and not Hispanic

____ (Items A-E should add to 100%)
100%

5. Does your school qualify for ESEA Title I assistance?

____ Yes - If Yes, approximately what number of students qualify for and what number of students are receiving ESEA Title I assistance?

____ Approximate number of students qualifying for ESEA Title I assistance

____ Approximate number of students receiving ESEA Title I assistance.

____ No

THANK YOU FOR YOUR COOPERATION

Standard Background Information Form for 9-Year-Olds

1. Does your family get a newspaper regularly?

Yes ○ No ○ I don't know. ○

2. Does your family get any magazines regularly?

Yes ○ No ○ I don't know. ○

3. Are there more than 25 books in your home?

Yes ○ No ○ I don't know. ○

4. Is there an encyclopedia in your home?

Yes ○ No ○ I don't know. ○

5. How much school did your father complete?
(FILL IN THE ONE OVAL which bests shows how much school your father completed.)

○ Did not complete the 8th grade

○ Completed the 8th grade, but did not go to high school

○ Went to high school, but did not graduate from high school

○ Graduated from high school

○ Some education after graduation from high school

○ I don't know.

6. Did your father graduate from a college or university?

Yes ○ No ○ I don't know. ○

7. How much school did your mother complete?
(FILL IN THE ONE OVAL which best shows how much school your mother completed.)

- ○ Did not complete the 8th grade
- ○ Completed the 8th grade, but did not go to high school
- ○ Went to high school, but did not graduate from high school
- ○ Graduated from high school
- ○ Some education after graduation from high school
- ○ I don't know.

8. Did your mother graduate from a college or university?

Yes ○ No ○ I don't know. ○

Standard Background Information Form for 13-Year-Olds

1. How much school did your father complete?
(FILL IN THE ONE OVAL which best shows how much school your father completed.)

- ⬭ Did not complete the 8th grade
- ⬭ Completed the 8th grade, but did not go to high school
- ⬭ Went to high school, but did not graduate from high school
- ⬭ Graduated from high school
- ⬭ Some education after graduation from high school
- ⬭ I don't know.

2. Did your father graduate from a college or university?

Yes ⬭ No ⬭ I don't know. ⬭

3. How much school did your mother complete?
(FILL IN THE ONE OVAL which best shows how much school your mother completed.)

- ⬭ Did not complete the 8th grade
- ⬭ Completed the 8th grade, but did not go to high school
- ⬭ Went to high school, but did not graduate from high school
- ⬭ Graduated from high school
- ⬭ Some education after graduation from high school
- ⬭ I don't know.

4. Did your mother graduate from a college or university?

Yes ⬭ No ⬭ I don't know. ⬭

5. Where did you live on your ninth birthday?

○ In the United States (Please specify the state or territory.)

_____

○ Outside of the United States (Please specify the country.)

_____

○ I don't know.

6. Which of the following does your family have at home? (Fill in one oval on each line.)

	Have	Do not have
A. Newspaper received regularly	○	○
B. Magazines received regularly	○	○
C. More than 25 books	○	○
D. Encyclopedia	○	○
E. Dictionary	○	○
F. Record player	○	○
G. Tape recorder or cassette player	○	○
H. Typewriter	○	○
I. Vacuum cleaner	○	○
J. Electric dishwasher	○	○
K. Two or more cars or trucks that run	○	○
L. Hand calculator	○	○
M. Home computer	○	○

7. How much time did you spend on homework yesterday?

- ⬭ No homework was assigned
- ⬭ I had homework but didn't do it
- ⬭ Less than one hour
- ⬭ Between 1 and 2 hours
- ⬭ More than 2 hours

8. Which one of the following best describes your grades so far in school?

- ⬭ Mostly A (a numerical average of 90-100)
- ⬭ About half A and half B (85-89)
- ⬭ Mostly B (80-84)
- ⬭ About half B and half C (75-79)
- ⬭ Mostly C (70-74)
- ⬭ About half C and half D (65-69)
- ⬭ Mostly D (60-64)
- ⬭ Mostly below D (below 60)

9. How much television did you watch yesterday?

⬭ None	⬭ 2 hours	⬭ 5 hours
⬭ 1 hour or less	⬭ 3 hours	⬭ 6 hours or more
⬭ 1 hour	⬭ 4 hours	

10. How many brothers or sisters do you have who are <u>older</u> than you?

None	1	2	3	4	5	6 or more
⬭	⬭	⬭	⬭	⬭	⬭	⬭

11. How many brothers or sisters do you have who are <u>younger</u> than you?

None	1	2	3	4	5	6 or more
⬭	⬭	⬭	⬭	⬭	⬭	⬭

12. What is your racial background?

- ⬭ American Indian or Alaskan Native
- ⬭ Asian or Pacific Islander
- ⬭ Black
- ⬭ White
- ⬭ Other (Please specify) _____

13. Is English the language spoken most often in your home?

Yes ⬭ No ⬭

14. Is a language other than English spoken in your home?

Often ⬭ Sometimes ⬭ Never ⬭

15. How much schooling would your parents like you to get? Do you think they want you to:

	Yes	No
A. graduate from high school?	⬭	⬭
B. go to vocational, technical, trade or business school?	⬭	⬭
C. go to a two-year or junior college?	⬭	⬭
D. go to a four-year college or university?	⬭	⬭
E. go to a graduate or professional school after college?	⬭	⬭

16. If there were no obstacles, how much schooling would you like to get? Would you:

	Yes	No
A. graduate from high school?	⬭	⬭
B. go to vocational, technical, trade or business school?	⬭	⬭
C. go to a two-year or junior college?	⬭	⬭
D. go to a four-year college or university?	⬭	⬭
E. go to a graduate or professional school after college?	⬭	⬭

Standard Background Information Form for 17-Year-Olds

1. How much school did your father complete?
(FILL IN THE ONE OVAL which best shows how much school your father completed.)

Did not complete the 8th grade

Completed the 8th grade, but did not go to high school

Went to high school, but did not graduate from high school

Graduated from high school

Some education after graduation from high school

2. Did your father graduate from a college or university?

Yes No I don't know.

3. How much school did your mother complete?
(FILL IN THE ONE OVAL which best shows how much school your mother completed.)

Did not complete the 8th grade

Completed the 8th grade, but did not go to high school

Went to high school, but did not graduate from high school

Graduated from high school

Some education after graduation from high school

I don't know.

4. Did your mother graduate from a college or university?

Yes No I don't know.

5. Where did you live on your ninth birthday?

○ In the United States (Please specify the state or territory.)

_____

○ Outside of the United States (Please specify the country.)

_____

○ I don't know.

6. Where did you live on your thirteenth birthday?

○ In the United States (Please specify the state or territory.)

_____

○ Outside of the United States (Please specify the country.)

_____

○ I don't know.

7. Which of the following does your family have at home? (Fill in one oval on each line.)

	Have	Do not have
A. Newspaper received regularly	○	○
B. Magazines received regularly	○	○
C. More than 25 books	○	○
D. Encyclopedia	○	○
E. Dictionary	○	○
F. Record player	○	○
G. Tape recorder or cassette player	○	○
H. Typewriter	○	○
I. Vacuum cleaner	○	○
J. Electric dishwasher	○	○
K. Two or more cars or trucks that run	○	○
L. Hand calculator	○	○
M. Home computer	○	○

8. How much time did you spend on homework yesterday?

- ○ No homework was assigned
- ○ I had homework but didn't do it
- ○ Less than one hour
- ○ Between 1 and 2 hours
- ○ More than 2 hours

9. Which one of the following best describes your present high school program?

- ○ General
- ○ Academic or college preparatory
- ○ Vocational or technical

10. Which one of the following best describes your grades so far in school?

- ○ Mostly A (a numerical average of 90-100)
- ○ About half A and half B (85-89)
- ○ Mostly B (80-84)
- ○ About half B and half C (75-79)
- ○ Mostly C (70-74)
- ○ About half C and half D (65-69)
- ○ Mostly D (60-64)
- ○ Mostly below D (below 60)

11. How much television did you watch yesterday?

○ None	○ 2 hours	○ 5 hours
○ 1 hour or less	○ 3 hours	○ 6 hours or more
○ 1 hour	○ 4 hours	

12. How many brothers or sisters do you have who are older than you?

None ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 or more ○

13. How many brothers or sisters do you have who are younger than you?

None ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 or more ○

14. A. What is your racial background?

○ American Indian or Alaskan Native

○ Asian or Pacific Islander

○ Black

○ White

○ Other

B. Is your ethnic heritage Hispanic (such as Mexican, Puerto Rican, Cuban, Central or South American or other Spanish Culture or origin)?

Yes ○ No ○

15. Is English the language spoken most often in your home?

Yes ○ No ○

16. Is a language other than English spoken in your home?

Often ○ Sometimes ○ Never ○

17. How often has each of the following been used in the courses you are taking this year? (Fill in one oval on each line.)

	Never	Seldom	Fairly Often	Frequently
A. Listening to the teacher's lecture	⬭	⬭	⬭	⬭
B. Participating in student-centered discussions	⬭	⬭	⬭	⬭
C. Working on a project or in a laboratory	⬭	⬭	⬭	⬭
D. Writing essays, themes, poetry, stories	⬭	⬭	⬭	⬭
E. Going on field trips	⬭	⬭	⬭	⬭
F. Having individualized instruction (small groups or one-to-one with a teacher)	⬭	⬭	⬭	⬭
G. Using teaching machines or computer-assisted instruction	⬭	⬭	⬭	⬭
H. Watching television lectures	⬭	⬭	⬭	⬭
I. Studying from textbooks	⬭	⬭	⬭	⬭
J. Library or media center assignments	⬭	⬭	⬭	⬭

18. How much schooling would <u>your parents</u> like you to get? Do you think they want you to:

	Yes	No
A. graduate from high school?	⬭	⬭
B. go to vocational, technical, trade or business school?	⬭	⬭
C. go to a two-year or junior college?	⬭	⬭
D. go to a four-year college or university?	⬭	⬭
E. go to a graduate or professional school after college?	⬭	⬭

19. If there were no obstacles, how much schooling would you like to get? Would you:

	Yes	No
A. graduate from high school?	○	○
B. go to vocational, technical, trade or business school?	○	○
C. go to a two-year or junior college?	○	○
D. go to a four-year college or university?	○	○
E. go to a graduate or professional school after college?	○	○

APPENDIX C

SAMPLES OF RESPONDENT LEVEL BACKGROUND QUESTIONS RELATED TC SPECIFIC SUBJECT AREAS

This appendix contains samples of background questions which have been included as part of subject area assessments. These questions are developed by consultants and NAEP staff responsible for developing the objectives and exercises for each assessment. They are included to provide information about attitudes and instruction for particular subject areas, as well as to enhance the analysis potential of the performance data.

Background Information on Music Experiences for 17-Year-Olds

Listed below are several kinds of music activities. Indicate how many years you have participated in each activity in school. Be sure to fill in one oval in each box.

Including elementary and junior high school, how many years have you taken or participated in

	Have Never Taken	Less Than 1 Year	1 to 2 Years	3 to 4 Years	5 or More Years
A. General Music Class or Music Appreciation?	⬭	⬭	⬭	⬭	⬭
B. Choir, Chorus or Glee Club?	⬭	⬭	⬭	⬭	⬭
C. Band or Instrumental Music?	⬭	⬭	⬭	⬭	⬭
D. Orchestra?	⬭	⬭	⬭	⬭	⬭
E. Introduction to Music, Music Literature or Music History?	⬭	⬭	⬭	⬭	⬭
F. Music Theory Class or Music Composition?	⬭	⬭	⬭	⬭	⬭

Background Information on Mathematics Experiences for 17-Year-Olds

Which of the following mathematics courses have you studied? Fill in one oval on each line. (If you have not studied a particular course, fill in the oval under "Not Studied.")

	Studied 1 school year	Studied 1/2 school year	Studied less than 1/2 year	Not studied	I don't know.
A. General, Business or Consumer Mathematics	⬭	⬭	⬭	⬭	⬭
B. Introduction to Algebra (Pre-Algebra)	⬭	⬭	⬭	⬭	⬭
C. 1st year Algebra	⬭	⬭	⬭	⬭	⬭
D. 2nd year Algebra	⬭	⬭	⬭	⬭	⬭
E. Geometry	⬭	⬭	⬭	⬭	⬭
F. Trigonometry	⬭	⬭	⬭	⬭	⬭
G. Probability and Statistics	⬭	⬭	⬭	⬭	⬭
H. Computer Programming	⬭	⬭	⬭	⬭	⬭
I. Pre-Calculus/ Calculus	⬭	⬭	⬭	⬭	⬭

Background Information on Mathematics Experiences for 17-Year-Olds

For each of the following questions, fill in one oval in each box.

A. The metric system of measurement uses units like centimeters, liters, and kilograms. How often have you used the metric system of measurement?

Often	Seldom	Never	I don't know.
⬭	⬭	⬭	⬭

B. How often do you use a hand calculator?

Almost Daily	A few times a week	Less than once a week	Once a month	Never	I don't know.
⬭	⬭	⬭	⬭	⬭	⬭

C. Do you or your family own a hand calculator?

Yes	No	I don't know.
⬭	⬭	⬭

D. Does your school provide hand calculators for use in mathematics classes?

Yes	No	I don't know.
⬭	⬭	⬭

E. Does your school provide hand calculators for use in other classes?

Yes	No	I don't know.
⬭	⬭	⬭

Background Information on Writing Experiences for 17-Year-Olds

1. How many reports and essays have you written during the last six weeks as part of any school assignment? _____

2. In the general English, literature or grammar classes you have taken during the past two years, about what part of the class time was spent on instruction in how to write reports and essays?

   ○ None of the time

   ○ Little of the time

   ○ About one-third of the time

   ○ About one-half of the time

   ○ Most of the time

3. In addition to the general English, literature or grammar classes you have taken during the last two years, have you had or are you now taking any of the following courses concerned with now to write?

	Yes	No	I don't know.
A. Creative writing course	○	○	○
B. Remedial writing course	○	○	○
C. Other writing course (If other, please specify.) _____	○	○	○

4. Are you encouraged to jot down ideas and make notes about the topic of your paper before you write it?

Usually	Sometimes	Never	I haven't written any papers.
○	○	○	○

5. Are you encouraged to make outlines of your papers before you write them?

Usually	Sometimes	Never	I haven't written any papers.
○	○	○	○

6. Do you write a paper more than once before you turn it in to your teachers?

Usually	Sometimes	Never	I haven't written any papers.
○	○	○	○

7. When your papers are returned, do they have written suggestions on how to improve your writing?

Usually	Sometimes	Never	I haven't written any papers.
○	○	○	○

8. When your papers are returned, do your teachers discuss them with you?

Usually	Sometimes	Never	I haven't written any papers.
○	○	○	○

9. After your papers are returned, do you work on the paper again to improve it?

Usually	Sometimes	Never	I haven't written any papers.
○	○	○	○

10. Do you enjoy working on writing assignments?

Usually	Sometimes	Never	I haven't written any papers.
○	○	○	○

Background Information on Art Experiences for 9-Year-Olds

1. Outside of school, what kinds of art work do you do? Tell whether you do each of the following things.

   A. Outside of school, do you draw?

   ○ Yes

   ○ No

   ○ I don't know.

   B. Outside of school, do you paint?

   ○ Yes

   ○ No

   ○ I don't know.

   C. Outside of school, do you make collages by cutting and pasting paper, cloth and scrap materials?

   ○ Yes

   ○ No

   ○ I don't know.

   D. Outside of school, do you carve or make models with wood, stone, clay, metal or plastic?

   ○ Yes

   ○ No

   ○ I don't know.

2. The works of artists are shown in art museums and art galleries. How often have you visited art museums or art galleries?

- ⬭ Never
- ⬭ One time
- ⬭ About five times
- ⬭ About ten times
- ⬭ Fifteen or more times

Background Information on Art Experiences for 13- and 17-Year-Olds

Outside of school, what kinds of art work do you do? Tell whether you do each of the following things outside of school.

	Yes	No	I don't know.
A. Drawing	⬭	⬭	⬭
B. Painting	⬭	⬭	⬭
C. Making pictures by cutting and pasting paper, cloth and scrap materials	⬭	⬭	⬭
D. Carving or modeling in wood, stone, clay, metal or plastic	⬭	⬭	⬭
E. Print-making such as block printing, silk screening, etching	⬭	⬭	⬭
F. Making pottery, ceramics or mosaics	⬭	⬭	⬭
G. Weaving, macrame or knot-tying, or needlework such as embroidery, needlepoint, knitting, crocheting	⬭	⬭	⬭
H. Making photographs or films	⬭	⬭	⬭
I. Making jewelry	⬭	⬭	⬭
J. Creating designs or plans for things like clothes, toys, cars, houses, furniture	⬭	⬭	⬭

Background Information on Art Experiences for 17-Year-Olds

1. Please indicate in which of the following years you took art classes in school.

	Yes	No	I don't know.
Seventh grade	⬭	⬭	⬭
Eighth grade	⬭	⬭	⬭
Ninth grade	⬭	⬭	⬭
Tenth grade	⬭	⬭	⬭
Eleventh grade	⬭	⬭	⬭
Twelfth grade	⬭	⬭	⬭

2. Are you currently enrolled in an art class?

Yes	No	I don't know.
⬭	⬭	⬭

Background Information on Reading Experiences for 13- and 17-Year-Olds

1. How often do you read for your own enjoyment during your spare time?

   ○ Almost every day

   ○ Once or twice a week

   ○ Less than once a week

   ○ Never

2. Which one of the following statements best describes the kind of reading you do in your spare time?

   ○ I never read during my spare time.

   ○ I mostly read fiction; for instance, short stories, novels and plays.

   ○ I mostly read nonfiction; for instance, books and articles about famous people, places, history, current events and hobbies.

   ○ I read fiction and nonfiction about equally.

Background Information on
Reading Experiences for 9-, 13- and 17-Year-Olds

1. How important is it to be able to read?

   ○ It is very important.

   ○ It is somewhat important.

   ○ It is not important at all.

2. How much do you enjoy reading?

   ○ I enjoy it very much.

   ○ I enjoy it somewhat.

   ○ I do not enjoy it at all.

Background Information on
Reading Ability for 9-, 13- and 17-Year-Olds

What kind of reader do you think you are for your age?

- ○ A poor reader
- ○ A good reader
- ○ A very good reader
- ○ I don't know.

Background Information on
Citizenship/Social Studies Experiences for 17-Year-Olds

1. In each statement below, please fill in the oval that BEST describes the general conditions in your school.

   A. Students are encouraged to make up their own minds.

   ○ Always

   ○ Often

   ○ Sometimes

   ○ Rarely

   ○ Never

   B. Teachers try to get students to speak freely and openly in class.

   ○ Always

   ○ Often

   ○ Sometimes

   ○ Rarely

   ○ Never

   C. Students can feel free to disagree openly with their teachers.

   ○ Always

   ○ Often

   ○ Sometimes

   ○ Rarely

   ○ Never

D. Our teachers respect our opinions and encourage us to express them.

- ○ Always
- ○ Often
- ○ Sometimes
- ○ Rarely
- ○ Never

E. In some schools the students help decide about school affairs; in others, the teachers and administrators make almost all of the decisions. How is it in <u>your</u> school--do the students help decide about school affairs?

- ○ Always
- ○ Often
- ○ Sometimes
- ○ Rarely
- ○ Never

2. Civics, history or government courses pay enough attention to important political issues such as race relations, political demonstration, poverty, and issues of war and peace.

- ○ True about the courses
- ○ Somewhat true about the courses
- ○ NOT true about the courses

3. Civics, history or government courses give students an accurate picture of the way American politics works.

- ○ True about the courses
- ○ Somewhat true about the courses
- ○ NOT true about the courses

4. Civics, history or government courses don't provide students with new ideas about politics. The students already know much of what is taught.

   ○ True about the courses

   ○ Somewhat true about the courses

   ○ NOT true about the courses

5. Civics, history or government courses give students the knowledge they need to participate in politics.

   ○ True about the courses

   ○ Somewhat true about the courses

   ○ NOT true about the courses

6. In school, have you studied how to acquire information about political issues?

   ○ True about the courses

   ○ Somewhat true about the courses

   ○ NOT true about the courses

7. In school, have you studied how to analyze the values and alternatives involved in political issues?

   ○ A good deal

   ○ Some

   ○ Not at all

APPENDIX D

ADJUSTMENT OF RESPONDENT WEIGHTS BY SMOOTHING TO REDUCE RANDOM VARIABILITY OF ESTIMATED POPULATION PROPORTIONS

Background

As noted elsewhere, a weight is assigned to every individual who responds to an exercise administered in an assessment. The weight is the reciprocal of the probability of selection of the individual with adjustment for nonresponse, and the probabilities of selection are based on the estimated number of people in the target age populations; therefore, the weight for an individual estimates the number of similar people that that individual represents in the age population. The sum of the weights of all individuals at an age level responding to an exercise is an estimate of the total number of people in that age population in the year that the exercise was assessed. Similarly, the sum of weights for all individuals who took the exercise and who also are members of some demographic category (such as blacks) gives an estimate of the number of people in the age population, for the year, who are also members of the category. The ratio of the two totals estimates the proportional representation of the demographic category in the age population for the given year.

For each of the assessed age populations in each assessment year, separate estimates of the proportional representation of the various demographic subgroups are provided by each booklet administered to that age group in that year. Due to random sampling variability, the estimates of population proportions for a given year based on single booklets administered in the year will vary. In addition to whatever trends in population proportions over time that might exist, there is also random sampling variation in these proportions from year to year.

It is desirable to reduce the random variability of population proportions as much as possible, since this variability has an effect on performance estimates. For example, the percentage of acceptable responses for an age group is a function of the relative proportions of high-performing and low-performing groups. If the relative proportions of these groups are very different in different

assessments due to sampling variability, then a portion of the change in percentage of acceptable responses for an age group could be attributable to yearly sampling difference in the relative proportions of high- and low-achieving groups.

In addition to reporting performance estimates for an age group as a whole, National Assessment also reports performance for various subpopulations, such as whites or blacks. Because variability of subgroups within these subpopulations (such as males and females within the white subpopulation) influences the performance estimates for the subpopulations, it is desirable that fluctuations of proportions of all subgroups of each subpopulation be reduced as much as possible.

For each age and year, each of the various booklets administered will provide estimates of a given population proportion. Since these estimates are subject to booklet-to-booklet variability, a better estimate of the population proportion, which will have reduced variability, is obtained by combining the information from all booklets. However, these proportions vary from year to year due to random sampling variability or systematic differences in sampling procedures. An even better estimate of population proportions for any single year can be obtained by smoothing the proportions over several assessment years. The word "smoothing" is used here in the sense of fitting a smooth curve to a sequence of numbers by robust/resistant procedures. Smoothing estimates of population proportions reduces a large portion of the sampling variability while preserving, as far as possible, actual trends occurring in the age population.

After the population proportions have been smoothed, adjusted weights are derived for the assessed individuals so that the population proportions computed using the adjusted weights are equal to the smoothed proportions. The adjusted weights are then used for all analyses, such as estimation of performance.

Smoothing Procedures Used by National Assessment

The most direct way to smooth proportions is first to classify people into mutually exclusive multiway cells on the basis of their membership in categories of various important variables and then to smooth the proportions within each of the resulting multiway cells across years. Unfortunately, this procedure tends to produce a large

number of cells with few people and, consequently, quite unstable estimates of smoothed proportions.

To circumvent this difficulty, National Assessment has utilized various smoothing procedures since the 1976-77 assessment. Each of these procedures, which are all basically weighting-class adjustments applied independently to each age, is designed to control, to varying degrees, fluctuations in certain key subgroups while avoiding, as much as possible, instabilities due to small cells.

The procedure used for the 1976-77 assessment was a weighting-class adjustment applied independently to each age and reporting variable (nation, region, sex, and so on). The details of the procedure are given in Appendix B of technical report 08-S-21: Three Assessments of Science, 1969-77: Technical Summary (1979). While this procedure performs well, it is complicated and requires large amounts of time and computer resources to implement. By independently smoothing proportions within each reporting variable, it was possible to produce good estimates of the marginal proportions of people within each category of the variable while disturbing as little as possible the relationships between other reporting variables within the adjusted variable. However, this means that each individual had a different adjusted weight for each reporting variable under consideration. While this presents no problem for the estimation of performance within a reporting variable, the multiplicity of weights definitely complicates analysis, such as regression, that involve several variables.

Because of the complexity of the procedure used in 1976-77, a different and simpler procedure was adopted in 1977-78. This procedure is detailed in Appendix F of Report 09-MA-40, Procedural Handbook: 1977-78 Mathematics Assessment (1980). The 1977-78 smoothing procedure produced a single adjusted weight for each individual, and hence greatly reduced the complexity of subsequent analyses of performance data. The 1977-78 procedure involved applying a weighting-class adjustment independently to each age. The weighting classes, which were different at each age, consisted of individuals who were alike on certain demographic characteristics and who would be expected to have similar educational achievement characteristics. There were around seventy adjustment cells used for each age.

Although the 1977-78 procedure produced acceptable results, National Assessment in 1978-79 adopted yet another procedure that we believe has the best characteristics of the three procedures used. The new 1978-79 procedure, which is detailed below, has several advantages.

1. It produces a single adjusted weight for each individual.

2. It affords good control on the distribution of proportions of certain key variables.

3. It produces the greatest stability of performance estimates.

4. It is the easiest to implement.

Even though adjusted weights using this procedure differ from the corresponding adjusted weights from the other procedures, we intend to use weights obtained using this procedure for all future analyses of data assessed in earlier years. This is simply because we believe these weights to be the best available. We recommend that secondary users also adopt this philosophy.

The Current Smoothing Procedure

The first step in the 1978-79 smoothing procedure involved the partitioning of the population of age class eligibles into the six smoothing cells given in Exhibit 1. The same cells were used for all ages.

Exhibit 1. Smoothing Cells Used for the 1978-79 Procedure

Cell	Race	Region	Community Size (CS)
1	White	All	Big City +Fringe (BC +FR)
2	White	All	Medium City (MC)
3	White	All	Small Places (SP)
4	Black	SE	All
5	Black	Not SE	All
6	Other	All	All

Then, for each age and every year, the proportion of the population in each of the cells was estimated. For a given age and year, the proportion of the population in a particular cell was computed as the sum of weights of all respondents assessed in the given year who were of the specified age and who belonged in the cell, divided by the total of the weight of all respondents of the given age assessed in that year.

Each of the six cells was comprised of a sequence of estimated population proportions corresponding to the various years of assessment. Each such sequence of proportions was then smoothed by fitting robust/resistant lines. Using data from the U.S. Census and Current Population Surveys, trends in enrollment by age and race and by age and region were obtained. The data from these surveys were adjusted to correspond with NAEP definitions as much as possible. The resistant lines within the smoothing cells were constrained to satisfy the trends from the U.S. Census and Current Population Surveys data.

The final step in the smoothing procedure was to adjust the respondents' weights to be consistent with the smoothed proportions. Since each respondent takes only one booklet, the weight adjustments were done independently for each booklet. For a given age, year and booklet, population proportions using the original weights were obtained for each of the smoothing cells. Then the weights of all respondents of a given cell were multiplied by the ratio of the smoothed cell proportion to the proportion using the original weights. This produced the adjusted weights that are used in all analyses.

Adjustment of Weights by Users

The smoothed population proportions for 9-, 13- and 17-year-olds (in-school only) are given in Exhibits 2, 3 and 4, respectively. The columns of a table correspond to the smoothing cells while the rows correspond to the assessment year. For example, the smoothed population proportion of 9-year-olds in smoothing cell two (whites in medium cities) for 1972-73 is .1158.

Exhibit 2. Smoothed Frequencies from Ten Year Smooth by Smoothing Cell and Year for 9-Year-Olds

	CELL					
	1	2	3	4	5	6
RACE REGION CS	WHITE ALL BC+FR	WHITE ALL MC	WHITE ALL SP	BLACK SE ALL	BLACK NOT SE ALL	OTHER ALL ALL
YEAR						
69-70	0.3293	0.1258	0.3546	0.0535	0.0745	0.0624
70-71	0.3251	0.1224	0.3601	0.0547	0.0746	0.0631
71-72	0.3210	0.1191	0.3656	0.0558	0.0747	0.0639
72-73	0.3168	0.1158	0.3711	0.0570	0.0748	0.0646
73-74	0.3126	0.1124	0.3766	0.0581	0.0748	0.0654
74-75	0.3085	0.1091	0.3822	0.0592	0.0749	0.0661
75-76	0.3043	0.1058	0.3877	0.0604	0.0750	0.0668
76-77	0.3001	0.1024	0.3932	0.0615	0.0751	0.0676
77-78	0.2959	0.0991	0.3987	0.0627	0.0752	0.0683
78-79	0.2918	0.0958	0.4042	0.0638	0.0753	0.0691

Exhibit 3. Smoothed Frequencies from Ten Year Smooth by Smoothing Cell and Year for 13-Year-Olds

	CELL					
	1	2	3	4	5	6
RACE	WHITE	WHITE	WHITE	BLACK	BLACK	OTHER
REGION	ALL	ALL	ALL	SE	NOT SE	ALL
CS	BC+FR	MC	SP	ALL	ALL	ALL
YEAR						
69-70	0.3200	0.1309	0.3703	0.0513	0.0673	0.0602
70-71	0.3201	0.1256	0.3731	0.0521	0.0681	0.0610
71-72	0.3202	0.1202	0.3760	0.0528	0.0690	0.0618
72-73	0.3203	0.1149	0.3788	0.0536	0.0698	0.0626
73-74	0.3204	0.1096	0.3816	0.0544	0.0706	0.0634
74-75	0.3205	0.1043	0.3845	0.0552	0.0714	0.0642
75-76	0.3206	0.0989	0.3873	0.0560	0.0722	0.0650
76-77	0.3207	0.0936	0.3901	0.0567	0.0731	0.0658
77-78	0.3207	0.0883	0.3929	0.0575	0.0739	0.0666
78-79	0.3208	0.0830	0.3958	0.0583	0.0747	0.0674

Exhibit 4. Smoothed Frequencies from Ten Year Smooth by Smoothing Cell and Year for 17-Year-Olds

	CELL					
	1	2	3	4	5	6
RACE REGION CS	WHITE ALL BC+FR	WHITE ALL MC	WHITE ALL SP	BLACK SE ALL	BLACK NOT SE ALL	OTHER ALL ALL
YEAR						
69-70	0.3405	0.1447	0.3686	0.0415	0.0581	0.0466
70-71	0.3419	0.1386	0.3704	0.0427	0.0591	0.0472
71-72	0.3432	0.1326	0.3722	0.0440	0.0602	0.0478
72-73	0.3446	0.1265	0.3740	0.0452	0.0612	0.0484
73-74	0.3460	0.1204	0.3759	0.0465	0.0622	0.0491
74-75	0.3474	0.1143	0.3777	0.0477	0.0633	0.0497
75-76	0.3487	0.1082	0.3795	0.0490	0.0643	0.0503
76-77	0.3501	0.1021	0.3813	0.0502	0.0653	0.0509
77-78	0.3515	0.0961	0.3831	0.0515	0.0664	0.0515
78-79	0.3529	0.0900	0.3849	0.0527	0.0674	0.0522

To adjust respondent weights to be consistent with the smoothed proportions, the following procedure is followed:

1. For each booklet, classify the respondents according to smoothing cell and obtain the raw population proportions for each cell. For example, the raw proportion for a booklet of 9-year-olds in smoothing cell four is the total of the weights of all 9-year-olds in the booklet who are black and in the Southeastern region, divided by the total of the weights of all respondents to the booklet.

2. For each booklet and smoothing cell, obtain a weight adjustment factor as the ratio of the smoothed population proportion (for the appropriate age, year and smoothing cell) over the raw population proportion.

3. The adjusted weight for an individual is the product of that individual's original weight and the appropriate adjustment factor.

Changes in Smoothed Proportions as New Assessment Are Completed

Every time an assessment is completed, a new time point is added to each of the sequences of population proportions within the smoothing cells. This means that, even though robust/resistant procedures are used, the addition of a new point may somewhat change the values of smoothed proportions for prior years. Additionally, any changes in methodology will impact the estimates.

This means that the smoothed proportions, with the addition of the next assessment data, are apt to differ somewhat from the corresponding smoothed proportions without the new data. National Assessment has adopted the philosophy that the smoothed proportions, based on all currently available data using the best available algorithm, are the best available. Therefore, all subsequent analyses, for any year, will be done using this best available information, even though this may produce estimates that slightly differ from prior values (before implementation of the newest smoothed proportions).

APPENDIX E

NONRESPONSE IN ASSESSMENT SAMPLES

In addition to sampling variability, estimates of population values computed from sample surveys might be subject to random error and systematic bias. Systematic bias, or nonrandom error, might result from estimation procedures, errors inherent in measurement and data collection procedures, and nonresponse.

Since nonresponse rates at ages 9 and 13 are relatively small, the following discussion concerns 17-year-olds' response rates only.

Bias due to nonresponse is present in virtually every sample survey but is frequently ignored since it is difficult to estimate its size. A variety of factors contribute to nonresponse. Nonrespondents might be either difficult to notify or reluctant to particpate once they are notified; some might be absent from school during the entire contact period with item administrators. However, these nonrespondents can be important, since, if they respond differently than the people actually included in the sample, estimates of percentage based solely on the sample are biased and not properly representative of the age population being assessed.

To provide some information about the size of the bias due to nonresponse in National Assessment surveys, the Research Triangle Institute, Raleigh, North Carolina, was asked to conduct a special study of nonrespondents during the 1972-73 assessment of science and mathematics. The study was conducted on the age population of eligible 17-year-olds who, at the time of the assessment, were listed as enrolled in school. Some of these students, in fact, were no longer attending school at the time of the assessment. Eligibles had to be English-speaking, had to be physically and emotionally able to respond to exercises as administered and could not reside in an institution.

The results of the nonresponse study (Kalsbeek et al., 1975; Rogers et al., 1977) indicate that 17-year-olds listed as enrolled in school but not appearing at the designated time of assessment can be divided into two different groups. The first group of nonrespondents, which comprises about 80 percent of the total nonrespondent group, did not appear for the assessment because of conflicting school activities or illness. The performance of this group was not very different from the performance of students assessed during the regularly scheduled sessions. The second group of nonrespondents, which comprises about 20 percent of the nonrespondents, does not appear to be available in the schools at any time. They attend infrequently if they attend at all (for practical purposes, they have dropped out of school), or they have moved out of the school attendance area. In either case, these students should probably not have been listed in the in-school population of eligibles. This group, in contrast to the group of nonrespondents who were, in fact, attending school, performed more poorly on assessment questions than students assessed during the scheduled sessions. They actually performed at about the same level as dropouts.

The weights used by National Assessment to estimate the percentage of acceptable responses are adjusted for nonresponses. The adjustment assumes that the nonrespondents would perform, on the average, in a manner similar to those who did respond. However, the nonresponse study showed that the second group of nonrespondents, those enrolled in but not actually attending school, typically performed at a lower level than either those who did respond or the first group of nonrespondents. If the second group is included in the population of eligibles, the nonresponse adjustment procedure used by National Assessment would result in overestimates of the true percentages of acceptable responses.

Because the second group of students is effectively no longer attending school, it does not seem appropriate to include them in estimates for 17-year-olds in school. Thus, these students are not considered part of the population of eligibles and are excluded from the computations of percentage of the sample covered for 17-year-olds, in school, as shown in Table E-1.

Including the second group of students and then reducing bias due to their nonresponse would require the location and testing of some of these individuals. The difficulty and costs associated with supplementary data collection of the nonrespondents not actually attending

school are so great that this has not been a feasible alternative in recent years.

Table E-1 shows the average sample coverage per package (booklet) of exercises administered in 1977-78. The rate of coverage is based on an estimated total eligible age population of students who are available in school--for 17-year-olds, those enrolled minus the 20 percent of nonrespondents estimated to be enrolled but unavailable in school.

Table E-1. Number of Students Assessed and Percent of Sample Covered, by Age, and Assessment Year 1977-78

Age	Number of Packages	Total Number of Students Assessed	Average Number Assessed Per Package	Average Sample Coverage in Percent
9	7	17,190	2,456	87.2
13	11	26,661	2,424	85.2
17	12	26,757	2,230	73.2

NOTE: Computations for 17-year-olds do not include those enrolled in school but not actually attending. No special follow-up was conducted in 1977-78 due to budget limitations.

APPENDIX F

INFERENCE FROM SURVEY DATA AND ANALYSIS OF DATA FROM COMPLEX SAMPLE SURVEYS

NAEP is a sample survey that collects data by using a multistage design with unequal probabilities of selection of elements. Such a survey design is commonly referred to as "complex" sampling. It permits the collection of representative data for the population and many subgroups in an extremely cost effective manner. NAEP data are suited for various descriptive purposes.

1. Estimation of achievement by exercise and summary level for the age populations as a whole and for a variety of subgroups of the populations.

2. Estimation of changes in achievement on a variety of tasks (and summaries) over time. Change can be measured for populations and for subgroups.

3. Explorations of observed associations between variables of interest.

The various descriptive statistics should be computed taking the structure of the sample into account.

The NAEP data present certain difficulties in analysis, which are shared by all complex surveys. These difficulties include problems in inference due to the type of data and complications in analysis due to the sample design. These difficulties are detailed in the following sections.

Inference From Survey Data

Because the data collected from a survey are observational and do not result from the control of variables (as do data collected from a designed experiment) they require careful interpretation. In experimental research it is often possible to infer causation because the researcher has controlled the individual variables. In surveys, because the values of the various variables are not controlled and because the variables tend to be related in value to each other and to other extraneous variables, such

control is generally not possible. Because correlation, in itself, does not prove causation, the direction of causation is often unclear for many variables from survey data, such as behaviors and correlated attitudes.

In applying standard statistical tests, it is often the case that undesirable confounding occurs between pertinent variables and other, perhaps extraneous, variables; hence, tests performed in the presence of these, often unremovable, confounding variables require very careful interpretation. Because of this, NAEP data should be examined in an exploratory, rather than confirmatory, framework.

Complications in Analysis Due to Sample Design

Complications in analysis occur because of the multistage design of the NAEP sample. Many standard statistical procedures assume that data are acquired by means of a simple random sample of the population and that individuals are independent. Because the NAEP sample employs stratification, clustering and unequal probabilities of selection, these assumptions are not met.

Certain subgroups of the population are sampled at a higher rate than the remainder of the population in order to ensure adequate representation. Consequently, those subgroups, which tend to have different characteristics (including achievement) than the remainder of the population, are overrepresented in the sample. Analyses that ignore this are apt to produce biased and misleading information, since those groups may have unwarranted impact. This difficulty is avoided by conducting weighted analyses, in which the weight assigned to an individual is related to the reciprocal of his or her probability of selection.

Because of cost and administrative efficiency considerations, NAEP data are obtained by selecting a number of schools and then selecting a number of students within each of the schools. Since the students are selected in clusters, observations from various students are not independent. Student responses within a school tend to be relatively more homogeneous than student responses from different schools.

Ignoring the effect of stratificiation and clustering in analysis tends to produce severe underestimates of the variablity of statistics. Many studies (such as Ross, 1976; Kish and Frankel, 1974; Frankel, 1971) have demonstrated large influences of complex survey designs on sampling errors of various statistics, such as regression and correlation coefficients. It has been demonstrated (Shah, Holt and Folsom, 1977) that regression analyses of data from complex survey samples produce tests of significance that are generally too liberal when the structure of the sample is not taken into account. Additionally, as noted by Fellegi (1979) in his consideration of goodness of fit tests, the distribution of certain statistics, as well as their dispersions, can be affected by the sample design.

Because of the nonlinearity of many of the statistics of interest, it is not currently possible to exactly account for the sample structure in analysis. However, several procedures exist that approximately do this. Among these are:

1. jackknifing--the procedure used by NAEP (detailed by Folsom, 1977);

2. balanced repeated replications (detailed by McCarthy, 1969); and

3. Taylor series approximation (detailed by Folsom, 1977).

Shah, Holt and Folsom (1977) give procedures for estimation and hypothesis testing of regression models for data from a complex sample survey using Taylor series approximations. A general procedure for obtaining approximate variances by Taylor series approximations is given by Woodruff and Causey (1976).

Fellegi (1979) and McCarthy (1969) give procedures for using balanced repeated replications to conduct goodness of fit tests.

Folsom (1977) gives the procedures used by NAEP in its analyses using jackknifing methodology.

A comparison of the performance of the three procedures for means, variances, correlations and regression coefficients is given by Frankel (1971) and Kish and Frankel (1974).

Approximately Accounting for Sample Design With Design Effects

It may be possible to approximately account for the effects of the sample design by using an inflation factor, the design effect, developed by Kish (1967) and extended by Kish and Frankel (1974). The design effect for a statistic is the ratio of the actual variance of the statistic (taking the sample design into account) over the variance assuming a simple random sample with the same number of elements. The design effect may be used to adjust error estimates based on simple random sampling assumptions so that the effect of the design is approximately accounted for. In practice, this is often accomplished by dividing the total sample size by the design effect and then using this effective sample size in the computation of errors. It must be kept in mind, however, that the value of the design effect depends on the variables considered in a particular analysis as well as the clustering effects occurring among sampled elements.

Based on empirical results and theoretical considerations, Kish and Frankel (1974) have developed several conjectures about design effects:

1. Generally, the design effects for complex statistics from complex samples are greater than one; hence, variances based on simple random sampling assumptions tend to be underestimates.

2. The design effects for complex statistics (such as regression coefficients) tend to be smaller than the corresponding design effects for means of the same variables. Hence, the latter estimates, which are more easily computed, tend to give overestimates of the design effects of complex statistics.

3. The design effects of complex statistics tend to resemble those of means; variables with a high design effect for the mean tend to also have high design effects for complex statistics involving those variables.

National Assessment has computed design effects for various types of statistics. A key statistic that has been examined is the estimated performance, P, of a subgroup of the population on an assessment exercise. This estimate, which is a weighted mean of the responses of individuals in the subgroup to the exercise, has a design effect of the form

$$deff(P) = Var\ (P)/(P(1-P)/N).$$

In the above, N is the total number of sampled individuals in the subgroup who responded to the exercise and Var(P) is the jackknifed variance of P (which takes the sample design into account). Upon examination of the distributions of design effects across exercises within reporting category (i.e., whites, disadvantaged urban, Northeast) for a variety of ages, subjects and years, two things were noted:

1. The various distributions were remarkably similar, although there were relatively more large design effects for the smaller, more clustered, subgroups such as disadvantaged urban.

2. The centers of all distributions were close to two, with the majority of values being less than two.

Because of this, a design effect of two was deemed a reasonable value to use for complex analyses. Based on Kish and Frankel's conjectures, this value should be an approximate upper bound for design effects of complex statistics. In a small scale regression analysis on NAEP data, the mean value of the design effects of regression coefficients was, in fact, two.

National Assessment has conducted regression analyses using design effects to account for the sample design and has found them adequate for exploratory purposes. The procedure used by National Assessment for regression analysis (called standard regression) using standard statistical packages (such as SPSS, BMDP and SAS) is as follows:

1. Every individual has a sampling weight that gives the relative importance of that individual in the population. These weights are necessary to provide proper contributions of the various subgroups in the estimates.

2. Each individual is assigned a scaled weight that is the person's sampling weight times the factor $f = (N/2)/WTOT$. In the formation of f, N is the sample size, N/2 is the effective sample size, and WTOT is the total of the sampling weights of all N individuals in the analysis. The scaled weights sum to the effective sample size and maintain the relative weightings of the individuals.

3. A weighted regression analysis is then conducted using scaled weights.

The estimates of the regression parameters using this procedure are identical to those obtained using the more complicated Taylor series and jackknifing procedures given by Folsom (1977). Although the estimated variances of the parameter estimates obtained using the above procedure differ from the more exact variance estimates (from Taylor series and jackknifing), they tend to be similar and appear adequate for exploratory purposes. However, significance tests using the standard regression approach need to be modified. The actual number of degrees of freedom for error is approximately the number of variance estimation replicates minus the number of strata rather than the effective sample size minus the number of parameters, the value given by the standard analysis. Therefore, significance should be assessed by comparison to critical values with the smaller error degrees of freedom.

It should also be noted that the estimated covariances between regression parameters using the standard regression can differ substantially from those estimated using Taylor series and jackknifing. This is because the joint distributions of subgroup performance and subgroup size are accounted for in the Taylor series and jackknifing procedures while they are not in the standard regression.

Tests of goodness of fit and independence can be performed on NAEP data by using the design effect approach. To accomplish this, the counts of individuals within various cells are replaced by sums of scaled weights where, again, the sum across all subjects of the scaled weights is the effective sample size. Apart from the substitution of scaled weight totals, the analyses proceed in the usual manner. Fellegi (1979) has empirically investigated this approach and found that although the tests are somewhat liberal this approach performs acceptably well for survey designs such as National Assessment's.

APPENDIX G

ANALYSIS ADVISORY COMMITTEE (ANAC)

The ANAC membership from 1965 to the present is as follows:

John P. Gilbert, chairman 1977-80. Staff Statistician, Harvard University Computer Center (1970-80).

Frederick Mosteller, chairman 1973-76. Professor of Mathematical Statistics, Harvard University (1970-76).

John W. Tukey, chairman 1965-73. Associate Executive Director, Bell Laboratories, Inc., and Professor of Statistics, Princeton University (1965-).

Robert P. Abelson, Professor of Psychology, Yale University (1965-72).

Darrell Bock, Professor of Education, University of Chicago (1977-).

David R. Brillinger, Professor of Statistics, University of California at Berkeley (1973-).

William E. Coffman, Professor of Education, University of Iowa (1970-74).

Lee J. Cronbach, Professor of Psychology and Education, Stanford University (1965-69).

James A. Davis, Director of National Opinion Research Center, Chicago, Illinois (1973-74).

Janet Dixon Elashoff, associate chairperson 1980- . Research Statistician, Center for Ulcer Research and Education, University of California at Los Angeles (1973-).

Gene V Glass, Associate Chairperson 1980- . Professor of Education, University of Colorado (1975-).

Lyle V. Jones, Professor of Psychology, L. L. Thurstone Psychometric Laboratory, University of North Carolina (1965-).

Lincoln E. Moses, Professor of Statistics, Dean of the Graduate School, Stanford University (1974-77).

Ralph W. Tyler, Science Research Associates, Chicago, Illinois (1968-69).

Lloyd Bond, Research Associate, Learning Research and Development Center, University of Pittsburgh, Pittsburgh, Pennsylvania (1980-).

BIBLIOGRAPHY

Access to School Districts, Schools and Nonstudents, no. 12-IP-56. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1980.

Contributions of the National Assessment to Understanding the Problems of Literacy and Equity, no. 12-IP-52. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1980.

Fellegi, I. P. "Approximate Tests of Independence and Goodness of Fit Based on Stratified Multi-Stage Samples," Survey Methodology, vol. 4, 1979.

Folsom, Ralph E. National Assessment Approach to Sampling Error Estimation, Sampling Error Monograph 25U-796-5. Research Triangle Park, N.C.: Research Triangle Institute, N.C., 1977.

Frankel, Martin R. Inference from Survey Samples. University of Michigan, Ann Arbor, Mich.: Institute for Social Research, 1971.

Kalsbeek, W. D. et al. No Show Analysis, final report, Raleigh, N.C.: Research Triangle Institute, 1975.

Kish, Leslie. Survey Sampling. New York: John Wiley and Sons, 1967.

Kish, Leslie and Martin R. Frankel. "Inference from Complex Samples," Journal of the Royal Statistical Society Series B, vol. 36, 1974.

McCarthy, P. J. "Pseudo-Replications: Half Samples," Review of the International Statistical Institute, vol. 37, 1969.

Miller, R. G. Jr. "A Trustworthy Jackknife," Annals of Mathematical Statistics, no. 35, 1964.

Miller, R. G. Jr. "Jackknifing Variances," Annals of Mathematical Statistics, no. 39, 1968.

Miller, R. G. Jr. "The Jackknife -- A Review," Biometrika, no. 61, 1974.

Mosteller, F. and J. W. Tukey. "Data Analysis Including Statistics," in Handbook of Social Psychology, 2nd ed., edited by E. Aronson and G. Lindzey. Reading, Mass: Addison-Wesley, 1968.

Mosteller, F. and J. W. Tukey. Data Analysis and Regression (Chapter 8). Reading, Mass.: Addison-Wesley, 1977.

Procedural Handbook: 1977-78 Mathematics Assessment, Report 09-MA-40, 1977-78 Assessment. Denver, Colo: National Assessment of Educational Progress, Education Commission of the States, 1980. ISBN 0-89398-143-5.

Rogers, W. T. et al. "Assessment of Nonresponse Bias in Sample Surveys: An Example from National Assessment," Journal of Educational Measurement, vol. 14, no. 4, 1977.

Ross, Kenneth Norman. Searching for Uncertainty. Occasional Paper no. 9. Hawthorn, Victoria, Australia: Australian Council for Educational Research, 1976.

Science: Results by Race, Parental Education, Size and Type of Community, Report 7, 1969-70 Assessment. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1973. ERIC no. ED 081 638. ISBN 0-89398-278-4.

Shah, Babubhai V. et al. "Inferences about Regression Models from Sample Survey Data." Paper presented at the International Association of Survey Statisticians Third Annual Meeting, New Delhi, India, December 5-15, 1977.

The National Assessment Approach to Objectives and Exercise Development, no. 12-IP-55. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1980.

Three National Assessments of Science: Changes in Achievement, 1969-77, Report 08-S-00, 1969-70, 1972-73 and 1976-77 Assessments. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1978. ERIC no. ED 159 026. ISBN 0-89398-296-2.

Three Assessments of Science, 1969-77: Technical Summary, Report 08-S-21, 1969-70, 1972-73 and 1976-77 Assessments. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1979. ERIC no. ED 168 901. ISBN 0-89398-297-0.

Tukey, John W. Exploratory Date Analysis. Reading, Mass: Addison-Wesley, 1977.

Tukey, John W. "Some Considerations on Locators Apt for Some Squeezed-Tail (and Stretched-Tail) Parents." Paper prepared in connection with research at Princeton University, supported by the Army Research Office, summer 1975.

Woodruff, Ralph S. and Beverley D. Causey. "Computerized Method for Approximating the Variance of a Complicated Estimate," Journal of the American Statistical Association, vol. 71, 1976.